UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD Thesis

Lucas Chaves Lima

Semantic Classification and Evaluation

Advisors: Christina Lioma, Jakob Grue Simonsen, Maria Maistro

Handed in: May 31, 2021
This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

Abstract

This thesis presents a collection of research articles that make contributions in the area of semantic classification and evaluation. Semantic classification describes the automatic processing of data, such as text, by machines, with the goal of simulating "understanding" the intended semantics, and as a result of this making a decision, for instance about the topic being discussed, or how some text should be translated into another language, or whether some piece of information constitutes fake news. This area has seen tremendous development in recent years, especially with the wide spread use of artificial neural network architectures, practically leading to almost human-like performance. This thesis presents a series of contributions in the design of artificial neural network architectures that: 1) can capture with high accuracy the most salient parts of text, in terms of syntax, semantics and grammar; 2) can capture semantic compositionality accurately; and 3) that can accurately detect fake news using different types of supporting evidence. This thesis also presents a series of contributions in how text processing is evaluated. Specifically, this thesis presents: 1) a family of novel evaluation measures that can evaluate rankings with respect to several aspects, such as relevance, and credibility and usefulness; 2) the biggest to this day evaluation dataset for fake news classification; and 3) a method for improving the evaluation capacity of incomplete evaluation datasets. Collectively, the above contributions advance the state of the art in how machines process and understand text.

Resumé

Denne afhandling består af en samling forskningsartikler, der bidrager til forskning i semantisk klassificering og evaluering. Semantisk klassificering er automatisk behandling af data, såsom tekst, af maskiner, med det mål at simulere "forståelse" af den tilsigtede semantik, og som et resultat heraf, tager en beslutning, for eksempel om emnet, der diskuteres, om hvordan en tekst skal oversættes til et andet sprog, eller om hvorvidt information udgør falske nyheder. Dette område har oplevet en enorm udvikling i de senere år, især med den udbredte brug af kunstige neurale netværksarkitekturer, der næsten kan præstere samme ydelse som mennesker. Denne afhandling består af en række bidrag til designet af kunstige neurale netværksarkitekturer, som: 1) med stor nøjagtighed kan indfange de mest fremtrædende dele af tekst hvad angår syntaks, semantik og grammatik; 2) kan indfange semantisk sammensætning nøjagtigt; og 3) kan registrere falske nyheder ved hjælp af forskellige typer bevismateriale med stor nøjagtighed. Tillige indeholder afhandlingen også en række bidrag til, hvordan tekstbehandling evalueres. Specifikt indeholder afhandlingen: 1) en familie af nye evalueringsforanstaltninger, der kan evaluere rangeringer af søgeresultater under hensyntagen til flere aspekter såsom relevans, troværdighed og anvendelighed; 2) det til dato største evalueringsdatasæt til klassificering af falske nyheder; og 3) en metode til forbedring af evalueringskapaciteten for ufuldstændige evalueringsdatasæt. Samlet bidrager afhandlingen til at udvikle, hvordan maskiner behandler og forstår tekst.

Sommario

Questa tesi presenta una raccolta di articoli di ricerca che forniscono un contributo nell'area della classificazione semantica e della valutazione. La classificazione semantica descrive l'elaborazione automatica di dati, ad esempio in formato testuale, da parte di sistemi, con l'obiettivo di simulare "l'apprendimento" della suddetta semantica e con il risultato di prendere una decisione, ad esempio riguardo all'argomento di una discussione, o la traduzione di un testo in un'altra lingua, o se dell'informazione rappresenta disinformazione. Quest'area si è sviluppata enormemente negli ultimi anni, soprattutto grazie ad un uso diffuso di architetture con reti neurali artificiali, che hanno portato a prestazioni simili a quelle di una persona. Questa tesi presenta una serie di contributi riguardanti la progettazione di architetture neurali che: 1) possono individuare con accuratezza elevata le parti più salienti del testo, relativamente alla sintassi, semantica e grammatica; 2) possono interpretare in modo accurato la composizionalità semantica; e 3) possono riconoscere con accuratezza le notizie false usando tipologie diverse di prove a sostegno dei fatti. Inoltre questa tesi presenta una serie di contributi relativi all'elaborazione e valutazione del testo. Nello specifico, questa tesi presenta: 1) una nuova famiglia di misure di valutazione in grado di valutare una lista ordinata considerando vari aspetti come rilevanza, credibilità e utilità; 2) il più grande dataset, attualmente disponibile, per la classificazione di notizie false; e 3) un metodo per migliorare la valutazione con dataset incompleti. In generale, i contributi sopra menzionati fanno avanzare lo stato dell'arte in relazione a come i sistemi elaborano e interpretano il testo.

Περίληψη

Αυτή η διατριβή παρουσιάζει μια συλλογή ερευνητικών άρθρων που συμβάλλουν στον τομέα της σημασιολογικής ταξινόμησης και αξιολόγησης. Η σημασιολογική ταξινόμηση περιγράφει την αυτόματη επεξεργασία δεδομένων, όπως το κείμενο, από ηλεκτρονικούς υπολογιστές, με στόχο την προσομοίωση της "κατανόησης" της επιδιωκόμενης σημασιολογίας, και ως αποτέλεσμα, την αυτόματη λήψη απόφασης, για παράδειγμα σχετικά με το θέμα που συζητείται, ή πώς κάποιο κείμενο θα πρέπει να μεταφραστεί σε άλλη γλώσσα, ή εάν κάποια πληροφορία αποτελεί ψεύτικη είδηση. Αυτή η περιοχή έχει σημειώσει τεράστια ανάπτυξη τα τελευταία χρόνια, ειδικά με την ευρεία χρήση τεχνητών νευρονικών δικτύων, που ουσιαστικά οδηγούν σε σχεδόν ανθρώπινη απόδοση. Αυτή η διατριβή παρουσιάζει μια σειρά από συνεισφορές στο σχεδιασμό των τεχνητών νευρωνικών δικτύων που: 1) μπορούν να συλλάβουν με υψηλή ακρίβεια τα πιο εμφανή τμήματα του κειμένου, από την άποψη συντακτικού, σημασιολογίας και γραμματικής, 2) μπορούν να συλλάβουν με ακρίβεια τη σημασιολογική σύνθεση, και 3) που μπορούν να εντοπίσουν με ακρίβεια πλαστά νέα χρησιμοποιώντας διαφορετικούς τύπους αποδεικτικών στοιχείων. Αυτή η διατριβή παρουσιάζει επίσης μια σειρά από συνεισφορές στον τρόπο αξιολόγησης της επεξεργασίας κειμένου. Συγκεκριμένα, η παρούσα διατριβή παρουσιάζει: 1) μια οικογένεια νέων μέτρων αξιολόγησης που μπορούν να αξιολογήσουν την κατάταξη δεδομένων σε σχέση με διάφορες πτυχές, όπως η συνάφεια , η αξιοπιστία και η χρησιμότητα, 2) την μεγαλύτερη μέχρι σήμερα βάση δεδομένων αξιολόγησης για ψεύτικη ταξινόμηση ειδήσεων, και 3) μια μέθοδο για τη βελτίωση της ικανότητας αξιολόγησης των ατελών βάσεων δεδομένων αξιολόγησης. Συλλογικά, οι παραπάνω συνεισφορές προωθούν την τεχνολογία με την οποία οι ηλεκτρονικοί υπολογιστές επεξεργάζονται και κατανοούν το κείμενο.

Resumo

A classificação semântica, que descreve o processamento automático de dados, como em textos, tem o objetivo de simular a compreensão da semântica pretendida, o que resulta na tomada de decisão, por exemplo, em como algum texto deve ser traduzido para outro idioma ou se alguma informação constitui notícia falsa. Esta área teve um grande desenvolvimento nos últimos anos, especialmente com o uso generalizado de arquiteturas de redes neurais artificiais, praticamente, levando a um desempenho quase semelhante ao de um humano. Esta tese apresenta uma coleção de artigos resultantes de pesquisas que trouxeram contribuições na área de classificação semântica e avaliação. Artigos estes que, na área da semântica, contribuíram em projetos de arquiteturas de redes neurais artificiais que: (1) podem capturar, com alta precisão, as partes mais salientes do texto, em termos de sintaxe, semântica e gramática; (2) podem capturar, com precisão, a composicionalidade semântica; (3) podem detectar, com precisão, notícias falsas apoiando-se em diferentes tipos de evidências. No tocante a avaliação sobre o processamento de textos, os artigos apresentaram contribuições em diversos aspectos como: (1) a apresentação de uma família de novas métricas de avaliação, capaz de avaliar rankings com relação a vários aspectos, como relevância, credibilidade e utilidade; (2) o maior conjunto de dados de avaliação existente até o momento para classificação de notícias falsas; (3) um método para melhorar a capacidade de avaliação em conjuntos de dados incompletos. Coletivamente, as contribuições reveladas por estes artigos avançam o estado da arte em como as máquinas processam e entendem o texto.

Acknowledgements

First of all, I have to say that my whole Ph.D. journey was incredible! The academic knowledge acquired and my personal development is immeasurable. Of course, none of this would have been possible without the support of my family, friends, and colleagues.

I am incredibly grateful to Christina for her valuable guidance and encouragement, always providing her insightful observations, organization, and comments that led to this thesis. I am also profoundly grateful to Jakob for his priceless criticism and support, always presenting his sharp points and objections. I also thank Maria for her unquestionable technical quality and willingness to help. Sometimes she helped me, even formulating questions that are not clear in my mind – with extraordinary patience. It is impossible to measure how much I have learned from you all.

I would like to extend my gratitude to my office friends for good laughs daily. Especially, for my dear friends Casper, Christian, and Dongsheng, for their support. You were essential and have made this journey much more manageable! I would also like to thank my family and friends that were not directly involved but were always present in my life. I also would like to acknowledge the QUARTZ [1] project for not only making everything possible but also allowing me to meet such incredible professors and colleagues that were part of this project. Especially for Benyou and Guilherme, with whom I spent the most time talking about research and beyond.

The following is only in Portuguese because it just does not make any sense otherwise. Gostaria de expressar minha mais profunda admiração, apreço, respeito e amor a minha mãe (Carminha), pai (Amauri) e irmão (Higor). O seu apoio em todos os aspectos da minha vida, principalmente me ensinando a valorizar a educação, nunca desistir e continuamente me encorajando a atingir meus objetivos, foram primordiais em toda essa jornada.

Last but not least, plenty of thanks to my beloved wife Anne Elize for your love, support and for sticking with me in good and bad times through this journey. I love her and thank her for all these years and many more to come.

[1] https://www.quartz-itn.eu/

Contents

Abstract **i**

Resumé **ii**

Sommario **iii**

Περίληψη **iv**

Resumo **v**

Acknowledgements **vi**

I COMPREHENSIVE SUMMARY **1**

1 INTRODUCTION **2**

1.1 Preamble 2

1.2 Thesis Outline 3

1.3 Overview of Contributions of the Thesis 3

1.3.1 Advances in Semantic Classification 3

1.3.2 Advances in Evaluation 6

1.4 List of Publications 7

2 BACKGROUND **10**

2.1 Semantic classification 10

2.1.1 Semantic understanding in text 10

2.1.2 Text classification and fake news detection 13

2.2 Evaluation 15

2.3 Evaluation Measures 17

2.3.1 Evaluating evaluation metrics 19

2.4 Evaluation datasets 21
2.4.1 Datasets 21
2.4.2 Veracity Datasets 22
2.4.3 Handling test collection incompleteness 23

3 CONTRIBUTIONS **26**
3.1 Advances in Semantic Classification 26
3.1.1 Research Question 1: How to guide the learning process of automated end-to-end neural networks towards semantically, syntactically and grammatically salient information? 26
3.1.2 Research Question 2: How to detect compositional phrases when processing text? 27
3.1.3 Research Question 3: How to improve the automatic detection of fake news? 28
3.2 Advances in Evaluation 30
3.2.1 Research Question 4: How to evaluate rankings of documents with respect to several aspects in theoretically principled ways that are invariant to the number and type of aspects? 30
3.2.2 Research Question 5: How to create a benchmark for developing and evaluating multi-aspect methods of sorting data, specifically in the domain of fake news classification? 33
3.2.3 Research Question 6: Can we invent methods that are better at handling test collection incompleteness that takes into account the class imbalance? 34
3.3 Ongoing Work 35

4 CONCLUSIONS AND FUTURE WORK **37**
4.1 Summary of Conclusions 37
4.2 Directions for Future Research 39

II INCLUDED PUBLICATIONS **43**

5 Multi-head Self-attention with Role-Guided Masks **44**

6 Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings **53**

7 MultiFC A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims **61**

8 Automatic Fake News Detection: Are Models Learning to Reason? **75**

9 Principled Multi-Aspect Evaluation Measures of Rankings. **83**

10 Test collection incompleteness and unjudged documents **95**

Bibliography **101**

Part I

COMPREHENSIVE SUMMARY

1. *INTRODUCTION*

1.1 Preamble

This thesis presents a collection of research articles that make contributions in the area of semantic classification and evaluation. Semantic classification describes the automatic processing of data, such as text, by machines, with the goal of simulating "understanding" the intended semantics, and as a result of this making a decision, for instance about the topic being discussed, or how some text should be translated into another language, or whether some piece of information constitutes fake news. This area has seen tremendous development in recent years, especially with the wide spread use of artificial neural network architectures, practically leading to almost human-like performance.

This thesis presents a series of contributions in the design of artificial neural network architectures that:

1. Can capture with high accuracy the most salient parts of text, in terms of syntax, semantics and grammar;
2. Can capture semantic compositionality accurately;
3. Can accurately detect fake news using different types of supporting evidence.

This thesis also presents a series of contributions in how text processing is evaluated. Specifically, this thesis presents:

1. A family of novel evaluation measures that can evaluate rankings with respect to several aspects, such as relevance and credibility and usefulness;
2. The biggest to this day evaluation dataset for fake news classification;
3. A method for improving the evaluation capacity of incomplete evaluation datasets.

Collectively, the above contributions advance the state of the art in how machines process and understand text.

1.2 Thesis Outline

This thesis is composed of a selection of six research articles that the author of the thesis authored during the period of his Ph.D. studies, 2018-2021. The thesis is divided into two parts. The first part is a comprehensive summary, which is structured as follows. Chapter 2 presents the necessary background work on semantic classification and evaluation that is required in order for the reader to follow the remainder of the thesis. Chapter 3 presents a discussion of the contributions of each of the included papers and ongoing work that is not yet published. Chapter 4 concludes the thesis and proposes future research directions. The second part of the thesis contains the compilation of the original research papers.

1.3 Overview of Contributions of the Thesis

This thesis contributes advances to the area of semantic classification and evaluation.

1.3.1 Advances in Semantic Classification

The contributions to semantic classification that are included in this thesis can be grouped into three high level clusters, described below. The first type of contributions in semantic classification can be seen as addressing the research challenge of:

> *How to guide the learning process of automated end-to-end neural networks towards semantically, syntactically and grammatically salient information?*

State of the art in semantic classification is a type of artificial neural network architecture called Transformer structured around units called "attention heads". Although some recent studies have shown that some attention heads of the Transformer can intrinsically learn linguistically-interpretable roles [34, 158], others have shown that heads can be pruned without significantly impacting effectiveness (indicating redundancy), or even improving it (indicating potential errors contained in pruned heads) [158, 113]. Thus, guiding the heads to spread their attention on

salient information of the input is a promising direction to diminish errors and redundancy among the heads. To address this research challenge, we introduced a modified architecture called the Transformer-Guided-Attn, which explicitly guides the multi-head attention of the Transformer by using role-specific masks, such that different heads are designed to play different roles. First, we choose important roles based on findings from recent studies on interpretable Transformers. Then, in order to guide the learning in an end-to-end manner, we extend the self-attention mechanism by incorporating masks that will force the head to attend to specific parts of the input with respect to its role.

The second type of contributions can be seen as addressing the research challenge of:

How to detect compositional phrases when processing text?

Compositional phrases are multi-word phrases that contain at least 2 words, but whose meaning does not represent the "sum" of each of its individual words (e.g. red herring). Compositional phrases are difficult to handle due to their idiomaticity. Work has addressed the problem of compositional phrases, but none of the existing work addresses the fact that the compositionality of a phrase is not dichotomous or deterministic but instead varies across context. For instance, the phrase "heavy metal" can be compositional when it refers to metal as the material, but it can also be non-compositional, when it refers to the music genre. To address this research challenge, we extract evidence from the global context of where a multi-word phrase occurs and its local context (narrative where the phrase is used). Moreover, we enrich the global and local phrase contexts extracted from a large corpus with phrase information extracted from external knowledge bases. Then, we combine both the global and local context to enrich the representation of the phrase. We reason that, even though the global and local contexts cover several usage scenarios, knowledge bases can offer a piece of more comprehensive and declarative information about the phrase. In addition to a method to detect compositional phrases, we extend an existing benchmark dataset that consists of 1042 phrases that are noun-noun 2-term phrases [53] where we add to each phrase one or two usage scenarios (same phrase in different contexts) using crowd-sourcing assessors.

The third type of contributions can be seen as addressing the research challenge of:

How to improve the automatic detection of fake news?

Fake news detection models are primarily trained on annotated data, that is typically mined from human fact-checked websites [14, 63, 68, 70]. Since not all fact-checking websites use the same labels, the number of labels is large, and it is not always clear from the guidelines on fact-checking websites how they can be mapped onto one another. Thus, training a veracity prediction model is not entirely straightforward. We contribute a fact-checking model that, differently from existing models that employ a pipeline approach, learns to weigh evidence pages by their importance for veracity prediction. Moreover, it makes claim veracity prediction with disparate label spaces by implicitly learning how semantically close the labels are to one another in an end-to-end manner, handling the multiple labels from different sources. We also address this research challenge by investigating whether fact-checking models genuinely determine the veracity of a claim by learning to reason over evidence. Automated fact-checking models inference of the veracity of claim/websites is usually based on reasoning: given a claim with associated evidence, the models aim to estimate the claim veracity based on the supporting or refuting content within the evidence. Commonly, it is assumed that whenever a model increases the performance, the model is learning to reason about the relation between the evidence and the claim. However, one of the main issues of learning models is that they can often memorize artifacts and biases instead of genuinely learning. Thus, in addition to proposing new fact-checking models, a more fundamental problem is how to measure whether fact-checking models are truly learning to reason. To answer this research question, we contribute an investigation of whether the model learns to reason over the evidence by exploring the relationship and importance of both claim and evidence. We state that a model using both the claim and evidence should perform better on fact-checking than a model using only the claim or evidence. The underlying assumption is that, given only the claim as input to the fact-checking model, the model does not have enough information to determine the claim's veracity. Similarly, if the model is only given the evidence, it corresponds to being able to provide an answer to an unknown question. We find that utilizing only the evidence achieved the best performance for most cases, and encoding the claim together with the evidence was either negligible or harmful to the effectiveness. This highlights a significant problem related to what constitutes evidence in current approaches, and also questions how current models are being evaluated.

1.3.2 Advances in Evaluation

The contributions to evaluation that are included in this thesis can be grouped into three high level clusters, described below. The first type of contributions in evaluation can be seen as addressing the research challenge of:

> *How to evaluate rankings of documents with respect to several aspects in theoretically principled ways that are invariant to the number and type of aspects?*

To address this research challenge, we contribute a multi-aspect evaluation approach, called Total Order Multi-Aspect (TOMA). The rationale of this method is that given a ranked list, TOMA first defines a preferential order (formally *weak order relation*) between aspects and category labels (e.g., relevant and not relevant for relevance) and then aggregates across multiple aspects to obtain a list of single category labels that can be evaluated by any single-aspect evaluation measure. Precisely, we embed category label tuples in Euclidean space and we derive the weak order using bespoke distance functions (see Chapter 9). For example, when the items that are ranked are documents retrieved by a search engine, then weak order relation allows deeming documents "equally good" when it is impossible or undesirable to impose a strict total order, allowing to rank even those documents that are not comparable with respect to the partial order relation. This ensures, by definition, that the preference order among multi-aspect tuples of labels is not violated for any number and type of aspects. We map each multi-aspect label tuple to a single integer weight (this allows us to aggregate multi-aspect tuples of labels so that better tuples can be given greater weight). Lastly, having such a weak order and weight, we use any existing single-aspect ranking evaluation measure to assess the quality of the ranking and guarantee that the final measure score is theoretically well-defined.

The second type of contributions can be view as addressing the research challenge of:

> *How to create a benchmark for developing and evaluating multi-aspect methods of sorting data, specifically in the domain of fake news classification?*

As most data does not come with the aspect of interest (e.g., websites will not come with an assessment of if they are fake or not), automatic inference is necessary in many cases. One such case is the domain of fake news, which is the one we make

a contribution to. To learn to infer the veracity of claims/websites, we contribute the currently largest[1] fact-checking real-world dataset (see Chapter 7). Our dataset, uniquely among existing datasets, contains a large number of naturally occurring claims and rich additional meta-information. Precisely, our Multi-Domain Dataset for Evidence-Based Fact Checking of Claims (MultiFC) consists of 34,918 claims which were extracted from 26 fact-checking websites, along with evidence pages, the context in which they occurred, and rich metadata.

The third type of contributions can be seen as addressing the research challenge of:

> *How to develop methods to improve evaluation by handling test collection incompleteness while accounting for class imbalance?*

Although test collections are very important for training, evaluating and comparing retrieval systems or other applications, they are often incomplete: they tend to contain considerably fewer assessed than non-assessed documents. Within this minority class of assessed documents, considerably fewer documents are assessed as relevant than non-relevant to a query. To address the challenge of test collection incompleteness, we contribute a method to automatically infer relevance assessments from document similarities to complete the test collection while still accounting for the imbalance between relevant and non-relevant documents (see Chapter 10). Our method estimates the relevance label using the inter-document similarities between the unjudged document and a set of assessed documents. Furthermore, we address the problem that the discriminative signal of relevant documents (minority) is weakened by the much stronger (majority) noise signal of non-relevant documents. To overcome the imbalance between relevant and non-relevant documents, we use a tunable threshold parameter to only let judged documents influence the inference if their similarity to the unjudged document exceeds the threshold parameter.

1.4 List of Publications

The following publications are included as chapters in this thesis.

Chapter 5 Dongsheng Wang, Casper Hansen, Lucas Chaves Lima, Christian Hansen, Maria Maistro, Jakob Grue Simonsen, Christina Lioma. Multi-head Self-attention with Role-Guided Masks. In: Hiemstra D., Moens MF., Mothe

[1] at the time of publication of the research article [15]

J., Perego R., Potthast M., Sebastiani F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12657. Springer, Cham. https://doi.org/10.1007/978-3-030-72240-1_45 [168]

Chapter 6 Dongsheng Wang, Qiuchi Li, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. 2019. Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings. In Companion Proceedings of The 2019 World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 317–323. DOI:https://doi.org/10.1145/3308560.3316584 [169]

Chapter 7 Isabelle Augenstein, Christina Lioma,Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-basedfact checking of claims. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4685–4697, Hong Kong, China, 2019. Association for Computational Linguistics.

Chapter 8 Casper Hansen, Christian Hansen, Lucas Chaves Lima. Automatic Fake News Detection: Are Models Learning to Reason? ACL-IJCNLP 2021: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (To Appear). [64].

Chapter 9 Maria Maistro, Lucas Chaves Lima, Jakob Grue Simonsen, Christina Lioma. Principled Multi-Aspect Evaluation Measures of Rankings (Under revision). 2021, 30th ACM International Conference on Information and Knowledge Management (CIKM2021) [111]

Chapter 10 Lucas Chaves Lima, Casper Hansen, Christian Hansen, Maria Maistro, Jakob Grue Simonsen, Christian Lioma. Test collection incompleteness and unjudged documents (Under Revision). 2021, 30th ACM International Conference on Information and Knowledge Management (CIKM2021) [93]

The following publications were written during the Ph.D. studies, but have not been included as chapters of this thesis. These publications are listed below in reverse chronological order:

[94] Lucas Chaves Lima, Casper Hansen, Christian Hansen, Dongsheng Wang, Maria Maistro, Birger Larsen, Jakob Grue Simonsen, Christina Lioma. Denmark's Participation in the Search Engine TREC COVID-19 Challenge: Lessons Learned about Searching for Precise Biomedical Scientific Information on COVID-19. In Ellen M. Voorhees and Angela Ellis, editors, Proceedings of the Twenty-Nine Text REtrieval Conference, TREC 2020, Gaithersburg, Maryland, USA, November 16-19, 2020, volume - of NIST Special Publication. National Institute of Standards and Technology (NIST), 2020.

[95] Lucas Chaves Lima, Dustin Wright, Isabelle Augenstein, Maria Maistro. University of Copenhagen Participation in TREC Health Misinformation Track 2020. In Ellen M. Voorhees and Angela Ellis, editors, Proceedings of the Twenty-Nine Text REtrieval Conference, TREC 2020, Gaithersburg, Maryland, USA, November 16-19, 2020, volume - of NIST Special Publication. National Institute of Standards and Technology (NIST), 2020.

2. *BACKGROUND*

This chapter presents the necessary background work on semantic classification and evaluation that is required in order for the reader to follow the remainder of the thesis.

2.1 Semantic classification

The semantic processing of text by machines has been a core object of study within computer science for more than 50 years now [26]. Numerous overviews of the area exist, see for instance chapter 2 of [98]. The core problem is how to approximate human semantic understanding by machines, so that we can have human-like inferences that are informed from this semantic understanding, such as classification of text by topic, text translation, or fake news detection in text. This is a very broad area, and in this thesis, the focus has been on two high-level directions. The first one is how to improve semantic understanding[1] per se, by guiding the learning process to salient parts of the text. The second one is how to improve the classification of text by topic and fake news detection in text. Next, we present a brief overview on these two directions.

2.1.1 Semantic understanding in text

There exist several methods for semantic understanding of the text: A most basic approach to represent text and extract features is using a bag of words (BOW) [62]. BOW encodes words using a one-hot encoding, where each word count is considered a feature. The primary assumption is that the words can be interpreted individually, not considering any connection between grammar, order, or how words relate to each other. The main advantage of representing text using bag of words

[1] Here, we discuss the semantic understanding of text by computers. Recently, there is also work on the semantic understanding of the output of computers by humans, see for instance [13, 14].

is that its concept is simple and intuitive. It is commonly used as the first step to identify and extract characteristics such as using the Term Frequency (TF) of a word in a piece of text, and it is usually weighted using the Term Frequency - Inverse Document Frequency (TF-IDF) to represent the informativeness of the word in the text [112]. BOW has some limitations such as its sparsity and absence of semantics. As the vocabulary of words increases, keeping track of all the words' identities leads to a high dimensional vector as large as the whole set of words in the whole collection (vocabulary).

Another way to represent text is using word embeddings [115, 124], which overcomes the problems of both sparsity and lack of context/semantics present in BOW representations [60, 115, 125]. In word embeddings, words are encoded using low dimensional vectors in such a way that this vector will represent the meaning of the word, which is captured according to its relation to other words [60]. To learn word vectors/representations, word embeddings use machine learning algorithms to learn real-valued or dense vector representation for each vocabulary term in a corpus. A well-known technique for learning such word embeddings is word2vec that uses neural network models to learn the word embedding using a collection of text [115]. Word2vec relies on two different algorithms, one that tries to predict the word given its surrounding words, namely continuous bag-of-words (CBOW), and another that tries to predict the surrounding context of a given word. A different word embedding is Global vectors (GloVe)[124]. The GloVe underlying assumption is that word-word co-occurrence probabilities can potentially encode some form of meaning.

Although word embeddings offer practical applications by representing words according to the context they occur, they have limitations. For instance, compositionality plays a vital role in word embeddings, and not handling non-compositional phrases, where the meaning of a multi-word phrase is not decomposed on the meaning of each word (e.g., heavy metal), can mislead and impact the quality of the word embedding [145]. Several existing approaches aim to identify non-compositional terms automatically. Some approaches assume that higher similarity between the constituents of a phrase and each of its components is indicative that the phrase is non-compositional [17, 86, 156]. Others estimate the similarity between a phrase and versions of the exact phrase, where the component words are perturbed using their synonym [87, 105]. Lioma et al. [99] represent the original phrase and its perturbations as ranked lists and measure their correlation or distance. Recently, increasing efforts attempt to use word embeddings and deep

artificial neural networks for compositionality detection [72, 139, 176]. Salehi et al. [139] also compute the similarity between each word represented in vector and the component word vectors using different representations [115, 121, 138]. Yazdani et al. [176] used neural networks to learn the semantic composition and recognize non-compositional phrases like those that stand out as outliers, assuming that the majority are compositional. Salehi et al. [137] use Wiktionary and utilize the definition, synonyms, and translations of Wiktionary to detect non-compositional components. Although the work mentioned above can be seen as a step forward for compositionality detection, they do not handle polysemy. For instance, if we consider the phrase "red herring", it can refer to a dried smoked fish which is compositional, or it could refer to some information intended to be misleading or distracting, which is non-compositional. We state that the compositionality of a phrase is not deterministic and can vary according to different contexts. Thus we contribute a method to consider contextual compositionality detection (see Chapter 6).

A more recent representation was proposed to overcome the lack of context of word embeddings. In contextual embedding, intuition is that a word may have a different meaning depending on its context. Thus, each word depends on its surrounding context and will not have a static vector representation. Instead its representation is generated dynamically according to its context [128, 49]. This contextual embedding is a step further towards models that can understand high-level concepts across many sentences. Several different models have been proposed to obtain such contextual embeddings. Peters et al. [128] developed Embeddings from Language Models) ElMo, which obtains its contextual embedding by training a model using a very large dataset to predict the next word (forward) or previous word (backward) using LSTM [75], which is later concatenated to encode the left and right contexts. The state-of-the-art in learning contextual embeddings is the Transformer model and its attention mechanisms [155]. Generative Pre-trained Transformer (GPT) and GPT2 use the Transformer model [155] trained on a large corpus and later fine-tuned in a supervised manner. Its learning process follows the assumption of how we read in English (left-to-right) in such a way that it only attends to its left context. A more powerful contextual embedding is the Bidirectional Encoder Representations from Transformers (BERT) ([49]). BERT also uses a Transformer encoder, but differently from GPT and GPT2, it attends to bidirectional contexts during pre-training. Several others extensions of BERT have been proposed: RoBERTa [106], ALBERT [89], SpanBERT [84], all of which have

architectures using transformers.

The Transformer was initially proposed as an encoder-decoder model. Its central part is the multi-head self-attention mechanism, which looks at the relationship between words within its context. Recent studies show that while some attention heads often learn essential (and even linguistically-interpretable) roles [34, 113, 158], other heads are less important and can be pruned without significantly impacting effectiveness [113, 158]. While the above-mentioned work investigated the roles of the attention heads, just a few have tried to guide the heads to attend some specific roles. For instance, Strubell et al. [147] trained the multi-head model with the first head attending to a single syntactic parent token, while the rest being regular attention heads. Sennrich and Haddow [144] used additional linguistic features such as sub-word tags, Part-of-speech (POS) tags as extra features into an attention encoder and decoder model in order to improve the model. Motivated by the above, instead of guiding a single head and only adding extra features to the input data, we present a method that guides the multi-attention heads to specific roles towards semantically, syntactically, and grammatically salient information (see Chapter 5). Next, we discuss some applications that use the above-presented methods of semantic understanding in practical applications.

2.1.2 Text classification and fake news detection

The purpose of fact checking is to make predictions of the veracity of claims. The veracity of a claim indicates the degree of being factual in terms of right or wrong. Recently, the task of fact-checking has been automated using machine learning and Natural Language Processing (NLP) to predict the veracity of claims. Automatic fact-checking is commonly framed as a textual classification problem, where given a claim, the models aim to estimate the claim veracity (e.g., classify whether a claim is true or false). Text classification consists of the process of predicting to which of a set of classes a new unseen textual input observation belongs, based on a training set of data containing observations whose classes are known. The classification is usually categorized according to the number of pre-defined classes, which are binary (e.g., False or True) or multi-class (False, Half-True, True). Each instance of the dataset is represented by a feature vector which is commonly obtained from the input text using the methods described in Section 2.1.1. Lastly, given a training dataset with labelled instances, the classifier approximates a function such that for a new unseen instance and its features, the classifier can predict a label from the set of pre-defined labels.

The type of dataset utilized affects how automatic fact-checking methods work. Methods only taking claims as input typically encode those with CNNs or RNNs [126, 171], and potentially encode metadata [171] in a similar way. Methods for small datasets often use hand-crafted features that are a mix of BOW and other lexical features, e.g., Linguistic Inquiry and Word Count (LIWC), and then use those as input to a Support Vector Machine (SVM) or Multilayer Perceptron (MLP) [18, 114, 126, 131]. More complex methods involve using deep learning approaches often trained on larger datasets, and account for evidence documents. These methods are contextual and non-contextual embeddings, which encode the claim and evidence using RNNs or Transformers aiming to identify the compatibility between claim and evidence [6, 51, 73, 129, 146]. In this thesis, we also contribute a new joint evidence ranking and claim veracity prediction model that learns to weigh evidence pages by their importance for veracity prediction in a end-to-end manner (see Chapter 7).

One of the main issues in machine learning, and also in fact-checking, is to guarantee that models will generalize well to new unseen data [1]. Recent research has shown that several NLP machine learning models can be fragile and spurious. For instance, simple and small changes in the training examples can lead the models to fail drastically [3, 20, 52, 83, 91, 92, 120, 175] while others have shown that the models often memorize artifacts and biases instead of truly learning [2, 61, 120]. One reason is the lack of careful consideration of how the models are evaluated, using evaluation measures and completely disregarding their known limitations. For instance, several measures rely on n-grams overlapping which has limitations such as not considering the meaning of the words, the order that words appear in the sentence. This can lead to a completely non-sense output with a "perfect" matching with the reference would result in a perfect score or even assign higher scores to system outputs than to human-authored texts [28, 50, 96, 123, 130]. Another reason for the lack of generalizability is due to biases that can exist in the dataset itself [57, 180]. This means that, even when the models achieve SOTA performance on their reported models, there is no guarantee that such results are truly learning and generalize well to new datasets (even for datasets from the same task). Hence, a deeper analysis of improvements in effectiveness and investigating what is truly being learned is necessary. Motivated by the above, we contribute an analysis of several fact-checking models and investigate whether such models are truly learning or memorising (see Chapter 8).

2.2 Evaluation

Evaluation is the process of utilizing empirical data to make judgments about the quality of the process being evaluated. Evaluation should be systematic and impartial, and it involves several steps and methodologies to make judgments of quality on particular criteria. Evaluation is crucial for assessing and improving different technologies, programs, models, etc [88]. In the context of IR, search engines are traditionally evaluated in terms of efficiency and effectiveness. Evaluating efficiency refers to the computational cost, such as the response time of the search system in every step of the whole retrieval process, including building the index [65, 66, 67, 69, 103], computational resources [179], retrieval [109]. Methods for evaluating the effectiveness of the retrieval results are crucial for further developments of search systems as they allow the measurement of how successfully the search system meets its objective of fulfilling the users' information needs [44]. Traditionally, the effectiveness of IR systems refers to the ability of the model in retrieving search results that satisfy the user need underlying the query. Search engine queries are commonly short [79] or are part of a more elaborate search task [178], and although they represent the users' need, queries are often ambiguous and can mislead the search engine to retrieve results that do not satisfy the users' information need. This raises the question of how do we evaluate IR systems?

There are two modes on how we analyze the effectiveness of IR systems: user-centric and system-centric. User-centric evaluation can include lab user studies with surveys, interviews, eye-tracking, etc., and log analysis with A/B testing experiments and interleaving. Even though those experiments often offer rich results, they are expensive, time-consuming, and hard to replicate [164]. A common approach for evaluating IR systems is system-centric. This involves using test collections to mimic a retrieval task, at the cost of the realism, since it only mimics the task and a possible user using the system [42, 164]. Test collections have three components: a corpus, a set of topics, and relevance assessments. A corpus consists of a collection of documents where the IR system will search for relevant information. The topics are a surrogate of the users' information needs, Usually they have a field which is "query" or "title" that represents the reformulation of the user's need as a query. The relevance assessments are judgments over topic-document pairs where a pair is assessed as either relevant or non-relevant (later extended to multi-graded) conditioned to whether it satisfies the information need

of the user or not [112]. The assessments of relevance are human-made therefore it is time-consuming and expensive, which can lead the test collections to have some limitations (see Section 2.4.3). On the other hand, once the test collection is built, it supports automated evaluation permitting development, contrasting, and optimization of several different search systems [44, 173]. In addition, it allows reproducibility favoring re-usable and comparative evaluations.

After constructing the test collection, the quality of the search systems can then be assessed via evaluation measures. Intuitively, the evaluation measures estimate a score of a ranked list of documents with their relevance assessments, i.e., a ranked list of documents should place documents labeled "relevant" before documents labeled "moderately/partially relevant" and "non-relevant" to receive a high score. Most evaluation measures estimate the effectiveness of the system by trying to mimic a user behavior using the IR system, e.g., starting inspecting documents from the top all the way down in the search results. More precisely, the evaluation measure is a function that receives as input an ordered vector of relevance assessments, and returns a single numeric score, summarizing the effectiveness of the vector [173]. Several evaluation measures have been proposed to evaluate search results such as Precision (P) and Recall (R) [43], Average Precision (AP) [24],discounted cumulative gain (DCG) [81], Normalized discounted cumulative gain (nDCG) [80], Expected Reciprocal Rank (ERR) [30], Rank-biased precision (RBP) [119]. Although the above evaluation measures are well known and studied in the literature, they measure the effectiveness of a ranking simply considering relevance, not covering different aspects of complex information needs. More precisely, relevance is known to be multidimensional [21, 47, 101, 142], including for instance, *Novelty, Reliability, Scope, Topicality, Understandability, Habit,* and *Interest*. In this thesis, we denote such different dimensions as *aspects* and we contribute a family of multiple aspect evaluation measures.

When there is only a single aspect the evaluation is straight-forward, since one can use the order of a given aspect to decide the quality of the ranked list. For instance, when only relevance is considered, we can assess the results of a search engine according to relevance labels, i.e., documents labeled "highly relevant" should be ranked before documents labeled "fairly/marginally relevant" and "non-relevant". The quality of the ranking is then assessed using evaluation measures, widely used evaluation measures are AP [24] or nDCG [80] (see Section 2.3). The problem arises when we wish to evaluate rankings by more than one aspect si-

multaneously[2]. The information need of the user is subjective, thus, assessing a document in terms of relevance is also subjective [100, 102]. As previously mentioned, many aspects have been identified to influence the relevance assessment, and it is estimated that there can easily be more than 20 attributes [19, 174]. While several measures for a single aspect evaluation (relevance) exist in the literature, just a few simultaneously combine these multiple aspects into the evaluation. This multiple-aspect scenario introduces the challenge of how to define evaluation measures that accommodate both an increasing number of aspects and a varied number of labels per aspect. Next, we discuss one of the components necessary to facilitate multi-aspect evaluation, namely evaluation measures.

2.3 Evaluation Measures

Evaluation measure is a staple of IR, with a plethora of literature and methods stretching back over decades. (P) and (R) represent the starting point for IR evaluation measures [43]. Recall is the proportion of relevant documents retrieved by the total number of existing relevant documents while Precision is the number of documents retrieved as relevant that were truly relevant to a query. Another evaluation measure combines P and R by taking the harmonic mean of both, namely F-measure, to provide a single value that evaluates the effectiveness of the system [16]. AP is defined upon P, which takes the average of the precision at the rank position where there is a relevant document, not each document. If you do this up to rank k, then it is $AP@k$. RBP [119], was proposed under the assumption that a user might not want to inspect documents at lower rankings and included a persistent parameter to define the probability that a user will continue inspecting documents lower in the ranking.

The above-presented measures use the notion that relevance is binary, meaning that a document is either relevant or not relevant. Later, binary relevance was extended to consider the relevance of documents in a graded scale of relevance (e.g., fairly relevant, partially relevant). In order to accommodate such graded relevance, researchers developed several evaluation measures. The first developed evaluation measure to consider graded relevance was DCG [81], which was later extended to its normalized form nDCG [82]. Intuitively, it computes the accumulated gain a user obtains by examining the top-ranking results up to a cutoff k, than it is nDCG@k, where this gain is top-weighted. Although AP, nDCG and

[2]This also holds across modalities, see for instance [59].

RBP consider the rank position, none of them consider the documents previously inspected, thus, Chapelle et al. [30] created the Expected Reciprocal Rank (ERR), which intuitively discounts the rank position of document accounting for the rank of previously seen relevant documents.

In addition to considering multi-graded relevance, relevance is also multidimensional [47, 142, 21], spanning across several dimensions. The concept of multiple dimensions were introduced as an attempt to cover the several criteria that could have influenced notion of relevance. Borlund [21] defined into five dimensions (Novelty, Reliability, Scope, Topicality, and Understandability), and later extended to seven, including Habit and Interest. Some evaluation measures explore the multiple dimensions of relevance, with some defined for specific contexts: relevance and novelty such as α-nDCG [35], Mean Average Precision Intent Aware (MAP-IA) [4] and Intent Aware Expected Reciprocal Ranking (IA-ERR) [4, 31, 35]; relevance, novelty and user effort as Normalized Cube Test (nCT) [151]; relevance, redundancy, and user effort as Rank-Biased Utility (RBU) [9]; relevance and understandability, such as Understandability-biased Rank-Biased Precision(uRBP) [183] and Multidimensional Measure (MM) [122]. Although the above evaluation measures are well known and studied in the literature, most of them measure the effectiveness of a ranking regarding relevance, disregarding the fact that there are many other aspects that can influence effectiveness. For instance, one could assess the information presented regarding the *credibility* of the content [97], others could evaluate the effectiveness of the information presented by measuring its *usefullness* concerning how this content helps a user to complete a task [178]. While several measures for a single aspect evaluation (relevance) exist in the literature, just a few simultaneously combine these multiple aspects into the evaluation.

Efforts to deal with multiple aspect evaluation can be split into two groups: (1) evaluate the aspects individually using any appropriate single-aspect evaluation measure (e.g., AP, nDCG), and then aggregate the scores into a single score over all aspects; or (2) evaluate all aspects at the same time using any appropriate multi-aspect evaluation measure [9, 104, 151]. Existing evaluation measures that consider aspects beyond relevance are either prone to serious theoretical anomalies, or are limited to specific types/numbers of aspects, such as Normalised Local Rank Error (NLRE), Normalised Global Rank Error (NGRE), Normalised Weighted Cumulative Score (nWCS), Convex Aggregation Measure (CAM), and the Weighted Harmonic Mean Aggregating Measure (WHAM) [104]. For instance, NLRE is

limited for two aspects, and its extension to a greater number of aspects returns the distribution of scores compressed towards 0, which prevents fair evaluations [97]. MM [122] is ill-defined for rankings that do not retrieve any relevant document. CAM [104] and MM can range in different intervals depending on the distribution of the aspects. uRBP needs to retrieve an infinite number of relevant and understandable documents to achieve a perfect score, even if the documents are not available.

Motivated by the above, we define a family of principled multiple aspect evaluation measures, TOMA (see Chapter 9). The rationale of TOMA is that given a ranked list, TOMA first defines a preferential order (formally *weak order relation*) between aspects and category labels and then aggregates across multiple aspects to obtain a list of single category labels that can be evaluated by any single-aspect evaluation measure. Next, we discuss how to evaluate new evaluation measures and compare them to existing ones.

2.3.1 Evaluating evaluation metrics

As different evaluation measures make decisions in different ways, assessing that the evaluation measure is capturing the effectiveness of the search system is necessary to not jump to wrong conclusions [25]. Evaluation measures can be evaluated in two complementary ways: theoretically, by comparing their properties to those of other known valid measures, and empirically, by comparing them to other known valid measures when assessing ranked lists of documents. In both cases, the comparison allows to reason about the limitations and improvements of the new evaluation measure.

Theoretical analysis: Evaluating the measures in a theoretical way usually consists of checking whether an evaluation measure satisfies a set of desiderata [132]. In this direction, Mofat et al. [117] showed that some evaluation measures are flawed in different ways (e.g., dismissal of the user experience as the user interacts with the ranking, suggesting that a user knows the number of relevant documents in the dataset, and so on), and advocated a few design goals that good metrics should follow. Ferrante et al. [56] state that any single-aspect evaluation measure should satisfy two properties: replacement and swap. Replacement states that if we replace an item in a ranking with a better one, then the measure score should not decrease. Swap states that if we swap a better item at the bottom of the ranking with a worse item at the top of the ranking the measure score should not decrease.

Moffat [116] defined seven proprieties and summarized a few well-known metrics analyzing whether they do or do not satisfy these proprieties. In particular, they defined the following properties: Boundedness, Monotonicity, Convergence, Top-weightedness, Localization, Completeness, and Realizability. Lioma et al. [104] derived a list of 8 evaluation desiderata that evaluation measures should follow and checked whether their evaluation measures satisfied these desiderata.

Empirical analysis: The empirical assessments of evaluation measures are commonly done in terms of correlation [54, 108, 149, 163], discriminative power [134], informativeness [10], stability [25], intuitiveness [136], and unanimity [5]. One of the most common approaches to evaluate evaluation measures is by using *correlation* coefficients such as Kendall's-τ and Spearman's ρ [140]. The interpretation of the results of the correlation analysis depends on how the correlation is used. For instance, while a new evaluation measure that strongly correlates to an existing one is likely to represent redundant information [172], a strong correlation between users preference and the new evaluation measure is desired. Tague et al. [150] use correlation measures to show that different precision-based evaluation metrics were highly correlated over different TREC collections. Chen et al. [33] have calculated the correlation between evaluation metrics and users satisfaction [148], mimicking practical users search experience. Others calculated the correlation between the results given by an evaluation measure over some set of search systems and the results of the users' preference on the same set of search systems [74, 141]. Turpin et al. [153] investigated the correlation between evaluation metrics and users' performance.

Another way to evaluate evaluation measures is based on the *informativeness* of an evaluation measure [10, 177]. The intuition is that, given a ranked list of documents and the evaluation measure score for that list, the maximum entropy method infers the distribution of relevant and not relevant documents in the ranking . Then, one can compare the inferred precision-recall curve against the actual precision recall-curve to determine the informativeness of the measure, i.e., their statistical ability to predict outcomes on held-out data. The *stability* test [25] assumes that there are multiple query reformulations associated with the same topic, and each system is run against all the queries; the stability of the evaluation measure is then approximated by empirically examining how the behavior of the evaluation measure changes across different query sets. The *intuitiveness* test [136] is

based on the idea that whenever two complex evaluation measures[3] disagree in selecting the best system out of two systems, simple measures (e.g. precision, recall) are used as a term of comparison. This method relies on the underlying assumption that the simple measures accurately represent the measurement of each aspect, thus, better alignment with simple measures results in more intuitive complex measures. The *unanimity* test by Albahem [5] uses the assumption that if a system A evaluated with simple measures is better over all aspects than another system B, meaning that all the simple measures agree over all aspects, this unanimity should be reflected by the complex measure. The *discriminative power* of an evaluation measure quantifies how good an evaluation measure is in guaranteeing which of two retrieval systems is the best. The discriminative power can be computed using the bootstrap sensitivity method [134], meaning that the higher the score, the more discriminative (i.e., the better) the approach.

Out of the empirical approaches, only correlation and discriminative power apply for multiple aspect evaluation measures, while the remaining is not applicable: the informativeness test [10] requires a precision recall-curve, which cannot be defined for multi-aspect evaluation; the intuitiveness test [136] requires simple single-aspect measures (e.g., precision, recall), which do not apply to multi-aspect evaluation; the unanimity test [5], which is defined for multi-aspect evaluation, requires that all the simple measures agree over all aspects which is extremely rare as the number of aspects increase. A detailed discussion of the results of our TOMA framework is presented in Chapter 9. Next, we discuss another component necessary for evaluation, i.e., evaluation datasets.

2.4 Evaluation datasets

2.4.1 Datasets

Different datasets can vary in size, structure, types of attributes, etc., and they usually relate to a particular subject. In the context of IR, effort from many initiatives such as TREC [71], NTCIR [85], CLEF [127], have developed large document collections, topics, and relevance assessments to serve as test-bed to improve different retrieval and classification models related to several tasks. The TREC 8-9 and TREC 2001-2004, have focused on tasks such as topic distillation and traditional ad-hoc retrieval, providing test collections with topics and relevance assess-

[3]By complex evaluation measures we mean either rank-based measures as in [136] or multiple aspect measures as in [5]

ments [159, 160, 161, 162, 165, 166]. Later, other initiatives tried to combine the different dimensions of the relevance of the previous years and proposed a task where the aim was to combine the multiple dimensions. In particular, TREC Web Track 2009-2014 [37, 38, 39, 40, 45, 46] promoted a task where the goal was to evaluate not only topical relevance of the ranking but also relating to the coverage and redundancy of the ranking regarding a query [41]. This allowed the development and evaluation of retrieval models that considered the diversity of the ranking into account, developing further the search systems. Moreover, datasets on aspects that go beyond relevance were proposed, where relevance was considered just one among several aspects that can impact the effectiveness of the search systems. Thus, in TREC Task Track 2015-2016 [157, 178], the organizers provided a test collection that embodies assessments of relevance and usefulness. Specifically, documents were assessed not only in terms of the relevance of a document to the query but also regarding its usefulness in completing the users' task underlying the query. This test collection allowed the development and evaluation of systems that helped the user to complete search tasks.

2.4.2 Veracity Datasets

Despite current technological advances in search engine technology, during the COVID-19 pandemic, the spread of misinformation led people to make wrong decisions having severe consequences on peoples' lives [78]. IR systems play an important role in serving information to users, but controlling the spread of misinformation via IR systems is difficult [32, 167]. In particular, the spread of non-credible/misinformation is constantly increasing [167], and it should not be underestimated [77]. This problem escalates in scenarios where the content is uncontrolled such as the web. Since users commonly use search engines to guide them through decisions, this represents a serious threat.

To combat this problem, many different datasets have been proposed to serve as a test-bed to estimate the veracity of potentially fake news/non-credible information. A first dataset was released by Schwarz et al. [143], which consists of a dataset with credibility assessments for webpages retrieved for queries of different topics. Initiatives such as the Decision Track 2019 [97], provided a test collection for health-related topics. Specifically, along with relevance assessments for documents, it also provided assessments of credibility and correctness in one of the largest document corpora of web content, the ClueWeb12. The Health Misinformation Track 2020 [36], released a test collection including assessments of

relevance, credibility, and correctness over documents for COVID-19 related topics, allowing the development and evaluation of IR systems able to rank *credible* and *correct* documents before not credible and incorrect documents.

As most data does not come with the aspect of interest, automatic inference is necessary in many cases. One such case is the domain of fake news as websites themselves will not say if they are fake or not. Many datasets related to fact-checking have been proposed testing the veracity of claims or aiming the prediction of fake news. These datasets mainly consist of claims obtained from different sources, including Wikipedia, and fact-checking websites such as politifact.com and snopes.com along with labels of veracity [126, 152, 171]. Although the previous datasets on fact-checking provide a good step towards the improvement of models to filter misinformation, most of them are either too small or are artificially generated. This motivated one of the contributions of this thesis which is to develop an evaluation dataset. Moreover, we build the currently largest naturally occurring claims dataset for fact-checking which contains evidence supporting the labels and rich additional meta-information (see Chapter 7).

2.4.3 Handling test collection incompleteness

Evaluating and comparing different retrieval systems is a core task in information retrieval; central to this task is the availability of test collections large enough to reflect the real-world data the retrieval systems will use. Building test collections with proper size is costly and time-consuming, prohibiting human annotation of all documents. A common approach to reduce costs consists of using a small sampled set of the whole collection judged for each topic, typically based on top documents from existing retrieval systems, called pooling. Several approaches to create the pool have been proposed [11, 48, 58, 107, 118]. As IR datasets scale up in size, the difficulty in ensuring a balanced distribution of assessed versus non-assessed documents increases, resulting in datasets with many more documents being non-assessed than assessed [182]. The presence of unjudged documents can lead to unfair or outright wrong evaluation scores and system comparisons [23]: for example, given a query and a ranked list of documents for that query–returned by some system–if an unjudged document occurs nearer to the top of the ranking than a second unjudged document, there is no canonically correct way of determining whether the system got the ranking right. Moreover, computing the measure requires choosing how to deal with documents whose relevance to the query is unknown, and different choices may lead to substantially different scores, even for

standard measures such as P and R, regardless of whether cutoffs (e.g., precision $@K$) are used or not.

A commonly used approach to deal with unjudged documents consists of mapping any unjudged labeled document to a not relevant label. Assuming that all of the unjudged documents are not relevant for a topic is a straightforward but strong assumption and may bias the evaluation by favoring systems that contribute more to the pooling [23]. A simple alternative approach is to ignore unjudged documents and perform the evaluation solely with judged documents, but this can bias the evaluation towards the recall of the systems since it ignores the documents' position. For example, a system placing unjudged documents near the top of a ranking might be judged equally good as one placing only relevant documents at the top [135].

A possibility is to assign as not relevant or completing ignore unjudged documents. An alternative is to *infer* document judgments automatically to ensure that all documents have judgments. The idea behind this approach is that–even if inferred judgments can be incorrect–having better-than-random-chance judgments will amend the imbalance problem between non-judged and judged documents. In general, this approach can yield fairer comparison between systems and enable supervised or semi-supervised learning algorithms to be trained on larger datasets of annotated documents. Roitero et al. [133] approximate the relevance of a document by using the search results of IR systems. Precisely, if a document is ranked close to the top of the ranking by several different systems, then that document is considered relevant. The intuition behind this method is the same as pooling which is commonly used to build the test collection. Using this method can be problematic since it might favor IR systems that participated in the pooling, leading to unfair evaluation. A second approach is to infer the precision/recall curve first and then use it to estimate the labels of non-assessed documents [12]. The main issue with this approach is that different distributions of relevant and non-relevant documents can generate the same precision/recall curve, meaning that the assessments estimated with this approach do not necessarily match the actual relevance of non-assessed documents. A third approach estimates the relevance label using the inter-document similarities between the unjudged document and a set of assessed documents [29]. A problem with this approach is that it does not handle the class imbalance between relevant and non-relevant documents. Specifically, for most queries, the number of non-relevant documents is vastly greater than relevant documents, whence the discriminative signal of relevant documents is weakened

by the stronger noise signal of the non-relevant documents.

Motivated by the above, we contribute our method that infers relevance assessments from document similarities to complete the test collection, i.e., balance the distribution of judged and unjudged documents, while still accounting for the class imbalance between relevant and non-relevant documents (see Chapter 10).

3. *CONTRIBUTIONS*

This section summarizes the research questions (RQs) related to each of the contributions of this thesis. We divide the contributions into two categories (i) Semantic Classification incorporating the research questions RQ1, RQ2 and RQ3; (ii) Advances in Evaluation incorporating RQ4, RQ5 and RQ6. Each formulated research question is accompanied by the findings of some of the articles of the author of this thesis.

3.1 Advances in Semantic Classification

3.1.1 Research Question 1: How to guide the learning process of automated end-to-end neural networks towards semantically, syntactically and grammatically salient information?

The state of the art in learning representations of words is the Transformer model and its attention mechanisms. Recent studies have shown that some attention heads of the Transformer can intrinsically learn linguistically interpretable roles [34, 158], others have shown that it can also lead to heads that can be pruned without significantly impacting (indicating redundancy) or even improving (indicating potential errors contained in pruned heads) effectiveness [158, 113]. Sennrich and Haddow [144] have showed the usefulness of incorporating linguistic features as input features into an attention encoder. Instead, we want to guide the learning process to attend to such linguistic features without disturbing the language model itself. Based on this, we propose the following research question:

(RQ1) How to guide the learning process of automated end-to-end neural networks towards semantically, syntactically and grammatically salient information?

Guiding the heads to spread their attention on salient information of the in-

put is a promising direction to diminish errors and redundancy among the heads. To address this research challenge, we introduced the Transformer-Guided-Attn, which explicitly guides the multi-head attention of the Transformer by using role-specific masks, such that different heads are designed to play different roles. First, we choose important roles based on findings from recent studies on interpretable Transformers. Specifically, we selected the following roles: specialized (rare words and separators), syntactic (dependency syntax and major relations), window (relative position) roles. Having decided the roles, we generate role masks used to constrain the attention head to some input parts. To do so, for each input sentence, we introduce a n-by-n matrix with all values being constrained (not used by the head). Then, for each token, we change its value to be considered depending on the masked role, e.g., a mask with the role of separators will only "activate" tokens that are separators, keeping the remaining restricted. Lastly, in order to guide the learning in an end-to-end manner, we extend the self-attention mechanism by incorporating masks that will force the head to attend to specific parts of the input concerning its role, i.e., for each attention head, we combine the original attention head with the n-by-n matrix with values that must be considered/ignored. Lastly, we concatenate all the attention heads resulting in the multi-guided attention head (see Chapter 5). We empirically show that on both text classification and machine translation on seven different datasets, our approach outperforms competitive attention-based, CNN, and RNN baselines.

3.1.2 Research Question 2: How to detect compositional phrases when processing text?

Automatic compositionality detection refers to identifying multi-word phrases such where the "sum" of the meaning of their words does not represent the whole meaning of the multi-word phrase. Although several works have addressed the problem of compositional phrases, none of the current work addresses the fact that the compositionality of a phrase is not dichotomous or deterministic, but instead varies across contexts. For instance, considering the phrase heavy metal, depending on the context, the phrase can be compositional (when referring to a metal), but it can also be non-compositional (referring to a music genre). Previous work has shown this property of compositionality theoretically [99], but no existing work has attempted to account for such property in automatic compositionality detection methods. Thus, motivated by the above, we propose the following research question:

(RQ2) How to detect compositional phrases when processing text?

To address this research challenge, we extract evidence from the global context of where a multi-word phrase occurs and its local context (narrative where the phrase is used). Specifically, to define the global context of a multi-word phrase, we extract the occurrence of all of its surroundings (window of size n immediately before and after) in a corpus and use all these windows to calculate the multi-word phrase distributional semantics across the whole corpus. To compute the local context, we first define a ranking of all the global contexts concerning the usage scenario (e.g., query, snippets) by computing the similarity between them (i.e., higher similarity closer to the top of the ranking). Next, we select the top K most similar context windows to the usage scenario. Then, we build the phrase's local usage scenario context representation by linearly combining both the global and the local usage scenario representations. Moreover, we enrich the global and local phrase contexts extracted from the corpus with phrase information extracted from external knowledge bases. We reason that even though the global and local contexts cover several usage scenarios, knowledge bases can offer a piece of more comprehensive and declarative information about the phrase. We experimentally show the usefulness of adopting the knowledge base combined with our contextual representation model (CRM) method. Specifically, we show that combining local, global, and knowledge bases increase the overall performance. In addition, we show that combining our approach CRM with RNN leads to an improvements over RNN. Moreover, in addition to a method to detect compositional phrases, we extend an existing benchmark dataset that consists of 1042 phrases that are noun-noun 2-term phrases [53] where we add to each phrase one or two scenarios (if possible), using crowd-sourcing assessors.

3.1.3 Research Question 3: How to improve the automatic detection of fake news?

Fake news detection models are primarily trained on annotated data, that is typically obtained from human fact-checked websites [14, 63, 68, 70]. Such fake news detection models usually infer the veracity of claim/websites based on reasoning, i.e., given a claim with associated evidence, the models aim to estimate the claim veracity based on the supporting or refuting content within the evidence. However, since not all fact-checking websites use the same claim labels, the number of labels is large, and it is not always clear from the guidelines on fact-checking

websites how they can be mapped onto one another, which makes the process of building fact-checking models not straight-forward. In addition, even when the fact-checking models achieve SOTA performance on their reported results, there is no guarantee that such models genuinely learn and generalize well to new datasets (even for datasets from the same task). Thus, developing new fact-checking models and measuring what is being learned by existing ones is challenging. Based on this, we propose the following research question:

(RQ3) How to improve the automatic detection of fake news?

As previously mentioned, using data collected from many different sources introduces a challenge when training a veracity prediction model. We contribute a joint evidence ranking and claim veracity prediction model that, differently from existing models that employ a pipeline approach, learns to weigh evidence pages by their importance for veracity prediction. Our inference model embeds the tokens from the claim and evidence into a joint space using an attention-weighted bidirectional LSTM [75]. To handle the disparate labels, inspired by multi-task learning, we use a label embedding layer that learns the relationship and how semantically close the labels are one to another. The label embedding layer allows the model to implicitly learn how semantically close the labels are to one another in an end-to-end manner, handling the multiple labels from different sources. We trained the proposed model in our dataset (see Section 3.2.2), and we show that encoding metadata and evidence pages is promising, improving the final performance of the inference model (see Chapter 7).

We also address this research challenge by evaluating whether fact-checking models genuinely determine the veracity of a claim by learning to reason over evidence. Specifically, we questioned the predictive power of the evidence itself and whether it assists the model in enabling better reasoning. We investigate whether the model learns to reason over the evidence by exploring the relationship and importance of both claim and evidence. To do so, we start from the statement: "A model using both the claim and evidence should perform better on the task of fact-checking compared to a model using only the claim or evidence". A scenario where giving only the claim as input to the fact-checking model produces the highest performance, suggests that the fact-checking model memorizes some particular signals and uses them when making the prediction. Specifically, as the prediction is based purely on claims seen during training, the fact-checking model might not have enough information to decide whether the claim is false or true. In-

stead, the model is exploiting some potential biases encountered in training. Similarly, whenever the fact-checking model only uses the evidence as input, the model predicts veracity based on an unknown question (claim). Thus, a fact-checking model that performs the best, when only the evidence is provided, indicates that the model is not reasoning over the evidence concerning the question, but simply picking strong signals from the evidence itself or the differences in the evidence obtained from claims with varying veracity. To investigate this, we empirically evaluate several representative fact-checking models of ranging complexity (from simple term-frequency-based representations to contextual embeddings), and we evaluate each model on various datasets and using different inputs (Claim only, Evidence only, and Claim+Evidence). Specifically, we select three different fact-checking models: a term-frequency-based Random Forest, a GloVe-based LSTM model, and a BERT-based model. As a benchmark dataset, we used a subset of the MultiFC dataset. This subset was generated by selecting the 2 largest political domains (PolitiFact and Snopes). Since the domains use different labels to represent the veracity of a claim, we manually mapped all of its labels to the same scale. This allows better comparison since we have comparable labels across the datasets. Comparing the results across the different fact-checking models and input types, we find that utilizing only the evidence achieved the best performance for most cases, and encoding the claim together with the evidence was either negligible or harmful to the effectiveness. This shows the existence of strong signals in the evidence itself. Therefore improvements associated with the evidence are not an indication of the model learning to reason over the evidence regarding the claim, but by simply learning to use a signal inherent in the evidence itself.

3.2 Advances in Evaluation

3.2.1 Research Question 4: How to evaluate rankings of documents with respect to several aspects in theoretically principled ways that are invariant to the number and type of aspects?

As discussed in Section 2.3, IR evaluation measures have traditionally focused on defining principled ways for assessing the relevance of a ranked list of documents concerning a query. There may be situations where relevance is simply enough. However, in most (common) retrieval situations, multiple aspects may satisfy the user's information needs better (whence effectiveness should be measured using

more than one aspect). We reason that the ranking must be assessed according to more than one aspect, e.g., relevance *and* correctness, where each aspect may have its own set of categories, e.g., {not relevant, partially relevant, highly relevant} for relevance, but {incorrect, correct} for correctness. Only a few evaluation measures were developed to consider aspects other than relevance (e.g., relevance, usefulness) using a single measure able to account for multiple aspects. However, none of the existing evaluation measures accommodate both an increasing number of aspects and various categories per aspect. Existing measures are either defined specifically for certain aspects and do not extend to other types/numbers of aspects, or lack a solid theoretical basis and can therefore end up in situations where they, e.g., assign scores greater than their maximum or assign the maximum score to a ranking where all documents are labeled with the lowest grade for all aspects (e.g., not relevant, not credible, etc.). Consequently, this leads us to the following research question:

(RQ4) How to evaluate rankings of documents with respect to several aspects in theoretically principled ways that are invariant to the number and type of aspects?

When evaluating using a single aspect (e.g. relevance), we can induce a weak order using the relevance labels, i.e., documents labelled "highly relevant" should be ranked before "fairly/marginally relevant" and "non-relevant" documents. However, when working in the multiple aspect scenario, this weak order is not straightforward because there are documents that are not comparable. The presence of not comparable documents implies that it is not possible to rank the documents univocally. For instance, consider two aspects with two category labels: relevance {not relevant, relevant} and correctness {incorrect, correct}. There is not unequivocally way to decide whether a document that is {relevant, incorrect} is better than a document {not relevant, correct}. Existing multiple aspect evaluation measures does not explicitly account for non comparable documents, which can make then range in different intervals. For instance, MM and CAM sort each aspect and compute the evaluation measure score independently, then later aggregate the evaluation measure scores. In the presence of non-comparable documents, in any way the documents are sorted, it will end up penalizing one aspect or the other. This means that there is no ranking that can achieve the perfect score and the upper bound of the evaluation measure is not known. Thus, to answer this research question, we reason that we first need to define how to sort tuples of labels (define a weak order

relation), weights them accordingly and compute the measure score. To this end, we present a multi-aspect evaluation approach, called Total Order Multi-Aspect (TOMA). To obtain the weak order relation, we embed the labels of each aspect in Euclidean space and derive the weak order using bespoke distance functions (e.g., Euclidean, Manhattan, and Chebyshev), that is the "distance order". The idea is to consider the maximum element of the set of category label tuples (best label) and let the ordering of the documents be induced by the "distance" of each category label tuple to the best label. The weak order relation allows deeming documents "equally good" when it is impossible or undesirable to impose a strict total order, allowing to rank even those documents that are not comparable. This ensures, by definition, that the preference order among category labels tuples is not violated for any number and type of aspects. The distance order determines a preference among multiple aspect documents and consequently how to sort those items (the smaller the distance from the best label, the better the category label tuple). However, the distance order does not specify how to aggregate category label tuple into a single number single number to compute an IR evaluation measure. To do so, we define a weight function and map each equivalence class (set of the tuple of labels) to a non-negative integer based on the distance order. Lastly, having such a weak order and weight, we use any existing single-aspect ranking evaluation measure to assess the quality of the ranking and guarantee that the final measure score is theoretically well-defined.

In short, we conclude that TOMA overcomes the most common anomalies of existing measures, and, empirically evaluations (on 342 rankings using up to 4 aspects and up to 5 categories per aspect) shows that TOMA is more discriminative, has correlations that are noticeable different between the Ranking of Systems (RoS) generated by TOMA and by current state-of-art multiple aspect evaluation measures which shows that they are not redundant. We experimentally find that TOMA instantiated with Manhattan distance selects as best the IR system that returns the lowest amount of low quality documents in the top ranks. Moreover, it is also the one associated with the highest quality of documents further down the ranking. The resulting reflections and lessons learned are discussed in Chapter 9.

3.2.2 Research Question 5: How to create a benchmark for developing and evaluating multi-aspect methods of sorting data, specifically in the domain of fake news classification?

Despite relevance, several aspects can affect the effectiveness of IR systems [36, 76, 97, 178]. In particular, a significant amount of effort has been focused on controlling the spread of misinformation. Researchers have recently started to view the filtering of fake news/fact-checking as a task that can be partially automated through the means of inference models. The development of inference models strongly depends on the availability of a benchmark dataset that will serve as a testbed to compare and evaluate the different models. Existing inference models either rely on small datasets consisting of naturally occurring claims or datasets consisting of artificially constructed ones. These claims are obtained from different sources, including Wikipedia and fact-checking websites such as politifact.com and snopes.com, along with labels of veracity [126, 152, 171]. While these datasets offer a valuable contribution to automated fact-checking, the lack of large datasets of real claims is what led us to our next research question:

(RQ5) How to create a benchmark for developing and evaluating multi-aspect methods of sorting data, specifically in the domain of fake news classification?

We contribute the currently largest fact-checking real-world dataset. Our dataset, uniquely among extant datasets, contains many real-life claims and rich additional meta-information. Our proposed dataset, called Multi-Domain Dataset for Evidence-Based Fact Checking of Claims (MultiFC), consists of 34,918 claims extracted from 26 fact-checking websites. We submitted each collected claim claim to a google search API and collected the top 10 most highly ranked search results as evidence pages. In addition to the evidence pages, to better understand the context in which the claim occurred, we conduct entity linking. We identify and collect from the claim entities such as people, places, organizations, and named entities. The detailed statistics of the dataset are discussed in Chapter 7.

3.2.3 Research Question 6: Can we invent methods that are better at handling test collection incompleteness that takes into account the class imbalance?

The availability of benchmark datasets large enough to reflect real-world data is necessary for developing IR systems. Relevance assessments are essential for training (i.e., developing) and comparing (i.e., evaluating) IR systems. There are two class imbalance problem in test collections. First, as test collections depend on the human annotation of documents, they commonly do not have a proper size, and thus, many documents in typical test collections are *unjudged*. Second, a further class imbalance problem occurs within this small set of assessed documents: many more documents are non-relevant than relevant to a query. Although this reflects the situation in the real world where typically many more documents are non-relevant to a query than relevant, this represents a clear case of an ML problem with an imbalanced dataset where the discriminative signal of relevant documents (minority) is weakened by the much stronger noise signal of non-relevant documents (majority). Existing approaches to handle unjudged documents can be grouped into three categories: ignore the unjudged documents, assign the unjudged documents as not relevant, and infer the relevance of the unjudged document. While the first two bypasses the problem, the latter does not handle the class imbalance between relevant and non-relevant documents. This motivates our next research question:

(RQ6) Can we invent methods that are better at handling test collection incompleteness that takes into account the class imbalance?

To answer this question, we first provide a thorough analysis of the effectiveness of applying several class imbalance handling approaches when applied to inferring relevance assessments from document similarities. On top of these class imbalance handling approaches, we contribute to reducing the noisy impact of highly dissimilar documents. Precisely, we use the state-of-art method by Carterette et al. [29] to infer relevance labels. Then we address its limitation wrt. class imbalance with our Document Similarity Thresholding (DST) method by defining a tunable threshold parameter to only let judged documents influence the inference if their similarity to the unjudged document exceeds the threshold parameter. We empirically evaluate our method against the standard approach of inferring relevance assessments from document similarities, using eight different sampling approaches (with and without DST), using three different representations of document semantics (TF.IDF, GLOVE word embeddings, and BERT word embeddings), and using

data from five different TREC tracks (248 queries and 256 submitted runs). We find that our approach is consistently the best in the tasks of (i) inferring relevance assessments of unjudged, (ii) predicting the ranking of competing IR systems using inferred relevance assessments, and (iii) predicting NDCG scores of rankings using inferred relevance assessments. The remaining results and lessons learned are discussed in Chapter 10.

3.3 Ongoing Work

The findings from RQ4 show that the TOMA framework effectively handles the multiple aspect evaluation by defining a distance order to obtain a weak order relation across all aspects. As the TOMA framework provides several different choices of instantiations in the embedding function, the distance function and the weight function, it also introduces some practical challenges in terms of necessary choices to instantiate the framework. The embedding function maps a nominal label to integers on a certain scale. TOMA in Maistro et al. [111] is instantiated with an embedding function that maps the different aspects to different ranges, but all labels were equispaced to the same step of size 1. It is important to consider the relationship between different aspects. The relation of aspects to each other can directly affect the final result of the distance function. Thus, exploring different embedding functions and distances to compute the distance order requires further investigation. Based on this, the following question arises:

> *Can we obtain a weak order relation in such a way that it is invariant to the arbitrary choices of the embedding function and distance function?*

This is an ongoing work that we are currently investigating. We are conducting an in-depth analysis of different embedding functions and how they impact TOMA when instantiated with distance functions. Specifically, we analyzed three different embedding functions:

1. Mapping all aspects in the same interval, where all labels are equispaced in the given range
2. Mapping aspects to different ranges, but all labels equispaced with the same step of the size of one

3. Mapping aspects to a range twice the size as the relevance range, and we do not use equispaced

In short, preliminary results show that changing the embedding mapping strongly impacts the final result of distance function instantiated with Chebyshev distance and mildly impacts the Euclidean and Manhattan distance as the correlation stays in the mid-high range across different embeddings functions. This shows that, indeed, choosing different embeddings affects the results of TOMA.

In light of these results, we propose a new way to obtain the weak order relation by applying the skyline operator [22], which was originally proposed to extend database management systems with queries able to account for multiple aspects. We call this skyline order. The skyline operator is "invariant" to the different embedding functions. Specifically, if any other document across aspects does not dominate (has greater or equal labels in all aspects, and greater label in at least one aspect) a document, it will be included in the skyline set. This means that the number of equivalence classes returned by the skyline order does not depend on the embedding function, but solely on the number of labels for each aspect. Even though the skyline operator is originally defined to identify just the skyline set, i.e. the set of label vectors that are not worse than any other label vectors in all the dimensions, we extend this definition to define an order relation among documents. To do so, we use the skyline set iteratively on the space of documents, thus defining the skyline levels. Each skyline level consists of documents globally better overall on all aspects than any other document in the levels below. After computing all skyline levels, we can use the weight function and map each skyline set (set of the tuple of labels) to a non-negative integer. In this way, TOMA using skyline order avoids the arbitrary choice of embedding functions and distance functions. Preliminary results show that utilizing the skyline order to obtain the weak order relation is effective.

4. *CONCLUSIONS AND FUTURE WORK*

4.1 Summary of Conclusions

This thesis addresses challenges related to semantic classification and evaluation. In terms of semantic classification, we tackle the challenges by proposing new models to: accurately learn representations by exploiting most salient parts of the text, capture semantic compositionality, and detect fake news. With regard to evaluation, we proposed a new multiple aspect evaluation measure, developed a evaluation dataset for fake news classification and proposed methods for improving evaluation datasets.

Improving semantic representations We presented two methods to improve the semantic representation and understanding of the text. First, we presented Transformer-Guided-Attn, which guides the attention head of the self-attention mechanisms using a role-specific mask. The underlying assumption is that guiding the heads to spread their attention on specific input parts reduces redundancy among the heads. We evaluated our method on two different NLP tasks, text classification and machine translation, on seven different datasets. We concluded that our approach obtain accuracy gains of up to 2.96% in text classification and 13% BLEU gains in machine translation tasks. We also developed a novel method for compositionality detection where the compositionality of a phrase is contextual rather than static. We show that enriching the word embeddings with local and global contexts outperforms state-of-the-art methods and that using knowledge bases can lead to notable performance improvements.

Bridge Between Evaluation and Application of Fact Checking We presented the largest real-world fact-checking dataset, consisting of 34,918 claims collected from 26 fact-checking websites, rich metadata, and ten retrieved evidence pages per claim. We analyzed several variants of fact-checking models, varying the use

of evidence, entities, and metadata to show the utility of our dataset. Through experimental analysis, we show that encoding the metadata and evidence pages improve the overall performance. This paves the way for future work in refining fact-checking models using our benchmark dataset. In addition, we contributed a new joint model for ranking evidence pages and predicting veracity that, among all other different model variants, performed the best in terms of Micro and Macro F1. In the domain-specific analysis, we concluded that even though for a few domains with small numbers of instances, the veracity prediction seems to be very easy (model achieving perfect Micro F1 and Macro F1) for the domains with the more considerable amount of claims, the veracity prediction is hard. Specifically, the overall best performance achieved by our model was Macro F1 equals 0.625 and Micro F1 0.492, which is far from perfect.

Insights on the Improvement of Automatic Detection of Fake News The most common approaches of automated fact-checking are based on models that use evidence to support the veracity decision of a claim. Whenever the fact-checking models improve its effectiveness, it is assumed that the model is learning to reason over some evidence to predict the veracity of a claim. This thesis investigated whether effectiveness improvements of fact-checking models are due to the models learning to reason over the evidence or using some existing bias in the evidence. After investigating models of varying complexity and evaluated on multiple datasets, we showed that for models, both training and testing within the same dataset or training in one dataset and testing on a different dataset achieves the highest performance for most of the cases when only the evidence is used, completely disregarding the claim for which the evidence was retrieved. These findings highlight a potential problem in the way that the evidence is currently collected/used. Finally, we conclude that this work paves the way for future work in refining alternatives to select and utilize evidence in automated fact-checking properly.

Multiple Aspect Evaluation Measures: Related to multiple aspects evaluation measures, we thoroughly described and validated our proposed framework TOMA. The TOMA framework uses distance functions to obtain a weak order relation that allows ranking even those documents that are not comparable and weight each tuple of labels to a weight. In this way, TOMA allows evaluation of the weak order using any single-aspect evaluation measure. We show that TOMA overcomes the limitations of state-of-the-art multiple aspect evaluation approaches: (i) it is

defined for any type and number of aspects and categories; (ii) it is well-defined, allowing rankings to reach the maximum score without exceeding it; (iii) and it is easier to interpret. We concluded that besides overcoming the limitations of existing state-of-art multiple evaluation measures, TOMA has better discriminative power, and is better at rewarding high-quality documents across the ranking than prior approaches to multi-aspect evaluation.

Improving Evaluation Datasets In addition to MultiFC, a new dataset for fact-checking, we presented a method to improve evaluation datasets. Specifically, we handled the problem of test collection incompleteness using Logistic Regression to infer labels of unjudged documents, as per [29], and we use 3 different document representations (TF-IDF, Glove and BERT); we define our method, called, DST. We tested different oversampling approaches to address the class imbalance problem between rel/not relevant documents. To the best of our knowledge, these have never been applied to the problem of inferring labels in IR. We experimentally showed that DST is effective (both without and with oversampling). Albeit being simple, our method yields a mean gain of $+17.28\%$ without oversampling and $+10.72\%$ with oversampling.

4.2 Directions for Future Research

In this section, we discuss some directions for future research, directly inspired from the results of this thesis.

Directions and further analysis of proposed Multiple Aspect Evaluation Measures Our framework TOMA can benefit from further investigations and evaluations of its effectiveness. Specifically, we have analyzed our TOMA framework using system validation. As previously discussed in Section 2.3.1, there exist different ways to evaluate an evaluation measure. A way of doing it is through user studies where we can evaluate the alignment of our current approach with real users' preference. In addition to user studies, one can use an in-depth investigation of the theoretical properties of TOMA using the existing axiomatic treatments of effectiveness for IR retrieval measures [7, 8, 27, 55, 110], where whenever the evaluation measure satisfies one of the axioms, we can guarantee some assumptions regarding the measured behavior and properties. In the light of further directions, one promising direction is to consider that some documents can harm the evalua-

tion score. For instance, when a document contains relevant, credible but incorrect information, the document should be seen as a threat and should be penalized.

Exploring dynamic feedback to weight aspects according to users' preference In this thesis, we have proposed multiple aspect evaluation measures that estimate the effectiveness of the ranking towards the users' needs. The users' interest should reflect the weight given for each aspect during evaluation, and the importance of each aspect might vary accordingly to the users' context or need. One promising direction here is to consider that users emphasize the same aspects in each search session [154], in which case the importance of each aspect can be predicted based on the users' dynamic feedback within a session. Specifically, in a session, which encompasses multiple queries, we believe that the importance of each aspect measure can be estimated by using the users' feedback (weight of aspects of the clicked documents) on the first query of a search session. Thus, to evaluate the subsequent queries of the same session, we use the aspect importance obtained by the users' feedback to update the weights of each aspect during evaluation.

Broader Impact of Multiple Aspect Evaluation Measures In this thesis, we presented a multiple aspect evaluation framework that can be used regardless of the concrete applications in where such rankings are used. In the broad context of machine learning, our TOMA framework directly impacts the evaluation, and hence the design, of loss functions used in the training of algorithms that learn to rank items, in a setting where multiple quality aspects of the items must be considered simultaneously. While ranking evaluation measures can be chosen for ethically spurious purposes (e.g., measuring rankings of job applicants where ethnicity is an aspect), neither the scientific problems we treat, nor the solutions we provide, are more likely to be usable for questionable purposes than for purposes benefiting society. Similarly, as with any ranking evaluation measure, the use of the methods described in this thesis could be used outside their intended setting to justify ranking decisions (say, in college admission) as being more "fair" than the traditional non-automated methods they supplant. We stress that TOMA is primarily intended for providing improved learning performance of algorithms learning to rank items, in scenarios where full automation is already the preferred method, e.g. in information retrieval settings.

On the Filtering of Evidence of Automated Fact-checking We showed how some bias in the evidence can impact the learning outcome of the fact-checking models, which can mislead the interpretation of the results. In light of fixing such biases, a direct approach is to filter out evidence that give away the labels (e.g., label appear in the evidence itself). A straightforward solution is to remove evidence that comes from domains of other fact-checkers websites. The reason is that there can exist claims that were checked by different fact-checkers, which would give away the veracity. Another possible direction is to consider a temporal cut based on the claim date. Specifically, removing evidence that happens after the claim date not only would remove potential evidence that gives away the veracity label but also would simulate a more realistic scenario where the models do not have access to future data. Lastly, as the evidence is currently assumed to be useful for fact-checking models, a user study investigating whether or not humans could determine the veracity of a claim based on the evidence provided could provide informative insights on the usefulness of such evidence.

On the Construction of Fair Datasets for Automated Fact-checking Even though the amount of automated fact-checking is widely increasing, its results should be carefully analyzed to identify whether what is being modeled is the actual task of fact-checking or some potential disruption on the data. While this data extracted from portals such as politifact.com or snopes.com offers valuable contributions to further automatic claim verification work, they can inadvertently risk the fairness of fact-checking model output due to skewed distributions in the data. This skewed distribution has been observed in different scenarios, where the predictions of the models encode social biases found in web corpora [181]. Specifically, some implicit correlations are not appropriate for real-world applications and can lead the output of the learning model to stereotype bias. These stereotypes are often thought to be the result of class imbalance in the training data. However, it is known that bias can occur without existing imbalances in the training set so the source of bias is traced to particular features in the model [90]. As some of those particular features reflect some property (e.g., personal attributes such as name, country), such bias in the features can lead to stereotypes. For instance, if we consider the distribution of labels concerning one feature (speaker) on a fact-checking dataset, the distribution of labels can be completely skewed towards one label, leading to stereotypes (e.g., Trump always lies). Thus, investigating and better understanding these potential imbalances, as well as constructing datasets that

aim to reduce such bias, e.g., having an equal number of false claims about Trump vs. True claims in the data, may improve the resulting effectiveness and the fairness of the output.

Part II

INCLUDED PUBLICATIONS

5. *Multi-head Self-attention with Role-Guided Masks*

Dongsheng Wang, Casper Hansen, Lucas Chaves Lima, Christian Hansen, Maria Maistro, Jakob Grue Simonsen, Christina Lioma. Multi-head Self-attention with Role-Guided Masks. In: Hiemstra D., Moens MF., Mothe J., Perego R., Potthast M., Sebastiani F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12657. Springer, Cham. https://doi.org/10.1007/978-3-030-72240-1_45 [168]

Multi-head Self-attention with Role-Guided Masks

Dongsheng Wang(✉), Casper Hansen, Lucas Chaves Lima, Christian Hansen, Maria Maistro, Jakob Grue Simonsen, and Christina Lioma

Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
{wang,c.hansen,lcl,chrh,mm,simonsen,c.lioma}@di.ku.dk

Abstract. The state of the art in learning meaningful semantic representations of words is the Transformer model and its attention mechanisms. Simply put, the attention mechanisms learn to attend to specific parts of the input dispensing recurrence and convolutions. While some of the learned attention heads have been found to play linguistically interpretable roles, they can be redundant or prone to errors. We propose a method to guide the attention heads towards roles identified in prior work as important. We do this by defining role-specific masks to constrain the heads to attend to specific parts of the input, such that different heads are designed to play different roles. Experiments on text classification and machine translation using 7 different datasets show that our method outperforms competitive attention-based, CNN, and RNN baselines.

Keywords: Self-attention · Transformer · Text classification

1 Introduction

The Transformer model has had great success in various tasks in Natural Language Processing (NLP). For instance, the state of the art is dominated by models such as BERT [5] and its extensions: RoBERTa [12], ALBERT [9], SpanBERT [8], SemBERT [24], and SciBERT [2], all of which are Transformer-based architectures. Due to this, recent studies have focused on developing approaches to understand how attention heads digest input texts, aiming to increase the interpretability of the model [4,14,20]. The findings of those analyses are aligned: while some attention heads of the Transformer often play linguistically interpretable roles [4,20], others are found to be less important and can be pruned without significantly impacting (indicating redundancy), or even improving (indicating potential errors contained in pruned heads), effectiveness [14,20].

While the above studies show that the effectiveness of the attention heads is, in part, derived from different head roles, only scant prior work analyze the impact of *explicitly* adopting roles for the multiple heads. Such an explicit guidance would force the heads to spread the attention on different parts of the input with the aim of reducing redundancy. This motivates the following research question: *What is the impact of explicitly guiding attention heads?*

D. Wang and C. Hansen—Equal contribution.

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 432–439, 2021.
https://doi.org/10.1007/978-3-030-72240-1_45

To answer this question, we define role-specific masks to guide the attention heads to attend to different parts of the input, such that different heads are designed to play different roles. We first choose important roles based on findings from recent studies on interpretable Transformers roles; then we produce masks with respect to those roles; and finally the masks are incorporated into self-attention heads to guide the attention computation. Experimental results on both text classification and machine translation on 7 different datasets show that our approach outperforms competitive attention-based, CNN, and RNN baselines.

2 Related Work

The Transformer [19] was originally proposed as an encoder-decoder model, but has also been used successfully for transfer learning tasks, especially after being pre-trained on massive amounts of unlabeled texts. At the heart of the transformer lies the notion of multi-head self-attention, where the attention of each head is computed as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, is the query, K is the key, V is the value, and d_k is the key dimension. The input to each head is a head-specific linear projection, and the Transformer uses multi-heads such that the attention for each head is concatenated for a single output.

Recently, efforts have been made to explore how the Transformer attends over different parts of the input texts [4,7,20]. Clark et al. [4] investigate each attention head's linguistic roles, and find that particular heads refer to specific aspects of syntax. Voita et al. [20] study the importance of the different heads using layer-wise relevance propagation (LRP) [6], and characterize them based on the role they perform. Furthermore, Voita et al. [20] find that not all heads are equally important and choose to prune the heads using a L_0 regularizer, finding that most of the non-pruned heads have specialized roles.

Scant prior work exists on guiding the attention heads to have a specific purpose. Strubell et al. [18] train the multi-head model with the first head attending to a single syntactic parent token, while the rest being regular attention heads. In contrast, we explore multiple more complex predefined roles grounded in head roles discovered in recent work. Sennrich and Haddow [15] incorporate linguistic features (e.g. sub-word tags, POS tags, etc.) as additional features into an attention encoder and decoder model for the task of machine translation, in order to enrich the model. In contrast, our method also makes use of linguistic features, but instead of enriching the input, we use these linguistic features to define the role-specific masks for guiding the attention heads.

3 Multi-head Attention with Guided Masks

We incorporate role-specific masks for self-attention heads, constraining them to attend to specific parts of the input. By doing this, we aim to reduce the redundancy between

the heads, and force the heads to have roles identified in previous work as important. Then, we adopt a weighted gate layer to aggregate the heads.

We first define the multi-head self-attention with role-specific masks in Sect. 3.1 followed by a description of each role in Sect. 3.2. We denote our final attention guided Transformer model as Transformer-Guided-Attn.

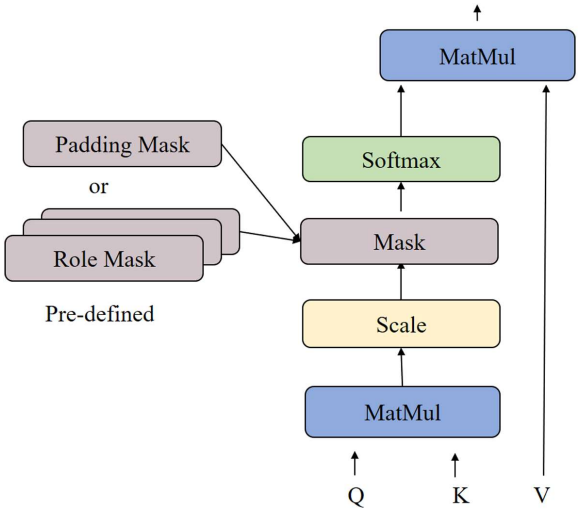


Fig. 1. Scaled-dot product with role mask or padding mask.

3.1 Multi-head Attention

We incorporate a role-specific mask into a masked attention head (mh) as:

$$\mathrm{mh}(\mathrm{Q},\mathrm{K},\mathrm{V},\mathrm{M_r}) = \mathrm{softmax}\left(\frac{\mathrm{QK^T}+\mathrm{M_r}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{2}$$

where M_r is a role-specific mask used to constrain the attention head. For an input of length n, M_r is an n-by-n matrix where each element is either $-\infty$ (ignore) or 0 (include). For multi-head self-attention, we introduce N role-specific masks for the first N heads out of a total of H heads ($N \leq H$). If N is strictly less than H, then the remaining heads are regular attention heads. Based on this, the multi-head attention can be expressed as:

$$\mathrm{MultiHead}(Q,K,V) = Concat(mh_1, mh_2, ..., mh_N, h_{N+1}, h_{N+2}, ..., h_H)W^O \tag{3}$$

where mh_i is the head with a role mask, and h_i is a regular head computed using Eq. (1).

A visualization of using the masks is shown in Fig. 1, where we associate the standard padding mask to regular attention heads. The padding masks ensure that inputs shorter than the model allowed length are padded to fit the model.

3.2 Mask Roles

We adopt the roles detected as important by Voita et al. [20] and Clark et al. [4]. We categorize them as 1) specialized (rare words and separators), 2) syntactic (dependency syntax and major relations), and 3) window (relative position) roles (see [10,11] for a linguistic basis of this categorisation). We include the separator role as Clark et al. [4] found that over half of BERT's attention, in layer 6–10, focus on separators. We describe these 5 specific roles below, which are used for creating role-specific masks.

Rare words (RareW). The rare words role refers to the least frequent tokens in a text. As defined by Voita et al. [20], we compute IDF (inversed document frequency) scores for all tokens and use the 10% least frequent tokens (highest 10% values according to IDF) in the sentence as the target attentions.

Separator (Seprat). The separator role guides the head to point to only separators. We extend the separator from $\{[SEP], [START], [END]\}$ to common punctuation of {comma, semicolon, dot, question mark, exclamation point}.

Dependency syntax (DepSyn). Dependency syntax role guides the head to attend to tokens with syntactic dependency relations. We assume this role can guide the head to attend to those–not adjacent–but still relevant tokens, complementary to the RelPos role (see below).

Major syntactic relations (MajRel). The major syntactic relations role guides the head to attend to the tokens involved with major syntactic relations. The four major relations defined by Voita et al. [20] are NSUBJ, DOBJ, AMOD, and ADVMOD.

Relative Position (RelPos). The relative position role guides the head to look at adjacent tokens, corresponding to scanning the text with a centered window of size 3.

For each role, we generate the guided mask for each input sentence by first producing an n-by-n matrix with all values as $-\infty$ (corresponding to ignoring all tokens initially). Then, we change the value of position (i, j) into 0.0, referring to the query token i with respect to the guided key token j, depending on the mask role.

4 Experiment

We experimentally compare our Transformer-Guided-Attn model to competitive baselines across 7 datasets in the tasks of text classification and machine translation. We make the source code publicly available on GitHub[1].

4.1 Classification Tasks

We consider two different classification tasks: sentiment analysis and topic classification. We compare our methods against six competitive baselines: the original Transformer [19]; multi-scale CNNs [22]; RNNs (BiLSTM) [3]; directional Self-attention (DiSAN) [17] that incorporates temporal order and multi-dimensional attention into the Transformer; phrase-level self-attention (PSAN) [23] which performs self-attention

[1] https://github.com/dswang2011/guided-attention-transformer.

across words inside a phrase; and Transformer-Complex-Order [21] that incorporates sequential order into the Transformer to capture ordered relationships between token positions. For the baselines implemented by us (marked in the Tables), we tune them as described in the original papers. For our Transformer-Guided-Attn, we consider a simple, but effective, way of selecting the combination of role-specific masks: For each layer, we fix 5 attention heads to be guided by the specific roles specified in Sect. 3.2, and let the remaining be regular heads. We tune the number of layers from $\{2,4,6,8\}$ and number of additional regular heads from $\{1,3\}$.

Dataset. The statistics of the datasets are shown in Table 1. We use the same splits as done by Wang et al. [21].

Results. As shown in Table 3, we observe consistent improvements compared to the best baseline for each dataset, except on MR where we perform as well as PSAN. Compared to the original Transformer model, we obtain accuracy gains of up to 2.96%, depending on the dataset, thus showing a notable performance impact from guiding the attention heads. Compared to DiSAN and PSAN, our proposed Transformer-Guided-Attn obtains consistent improvements over the original Transformer across all datasets, while DiSAN and PSAN both have lower performance for TREC and SUBJ (Table 2).

Table 1. Classification dataset statistics. CV means 10-fold cross validation.

Dataset	Train	Test	Task	Vocab.	Class
CR	4k	CV	Product review	6k	2
TREC	5.4k	0.5k	Question	10k	6
SUBJ	10k	CV	Subjectivity	21k	2
MPQA	11k	CV	Opinion polarity	6k	2
MR	11.9k	CV	Movie review	20k	2
SST	67k	2.2k	Movie review	18k	2

Table 2. Machine translation results. ⋆ marks scores reported from other papers.

Method	BLEU
Transformer [19]	34.3
AED + Linguistic [15] ⋆	28.4
AED + BPE [16] ⋆	34.2
Tensorized Transformer [13] ⋆	34.9
Transformer-Complex-Order [21] ⋆	35.8
Transformer-Guided-Attn (ours)	**38.8**

Table 3. Classification results (accuracy %). ⋆ marks scores reported from other papers.

Method	CR	TREC	SUBJ	MPAQ	MR	SST
Transformer [19]	82.0	91.8	93.2	88.6	77.7	81.8
Multi-scale CNNs [22]	81.2	93.1	93.3	89.1	77.8	80.9
BiLSTM [3]	82.6	92.4	93.6	88.9	78.4	81.1
DiSAN (Directional Self-Attention) [17]⋆	84.1	88.3	92.2	89.5	79.7	82.9
PSAN (phrase-level Self-Attention) [23]⋆	84.2	89.1	91.9	89.9	**80.0**	83.8
Transformer-Complex-Order [21]⋆	80.6	89.6	89.5	86.3	74.6	81.3
Transformer-Guided-Attn (ours)	**84.4**	**93.6**	**93.8**	**90.7**	**80.0**	**84.2**

4.2 Translation Task

We use the standard WMT 2016 English-German dataset [16] and use four baselines: Attentional encoder-decoder (AED) [15] with linguistic features including morphological, part-of-speech, and syntactic dependency labels as additional embedding space; AED with Byte-pair encoding (BPE) [16] subword segmentation for open-vocabulary translation; the tensorized Transformer [13]; and the Transformer-Complex-order [21]. The first two models are extensions on top of the basic AED [1]. For the models we implement, we follow the same tuning as in the classification experiments. We evaluate the machine translation performance using the Bilingual Evaluation Understudy (BLEU) measure.

Results. Our Transformer-Guided-Attn consistently outperforms the competitive baselines. Specifically, we observe gains of 8.2% compared to the best baseline, Transformer-Complex-Order, and close to 13% compared to the original Transformer. These gains are even larger than the results for the classification experiments, thus highlighting a significant performance impact from guiding the attention heads for the task of machine translation.

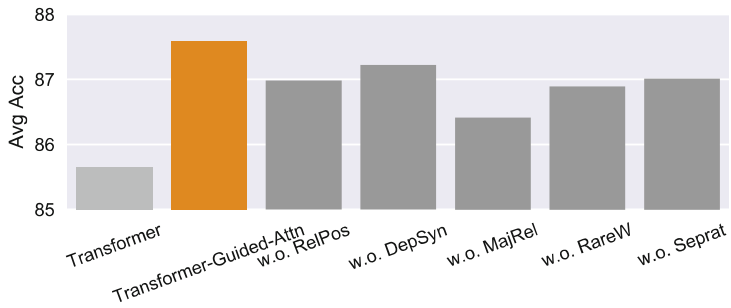


Fig. 2. Ablation study of Transformer-Guided-Attn when dropping each role individually.

4.3 Ablation Study

We now consider the performance impact associated with each role-specific mask. For each classification dataset, we run configurations of our Transformer-Guided-Attn with each role-specific mask excluded once and replaced with a default padding mask used in the Transformer. The average accuracy drop associated with excluding each role-specific mask is shown in Fig. 2, which also includes the average accuracy of the Transformer and our Transformer-Guided-Attn using all role-specific masks. We observe that the removal of each role has a negative impact on performance, where the major syntactic relations role (MajRel) has the largest impact. Thus, collectively all roles contribute to the performance of the full Transformer-Guided-Attn model.

5 Conclusion

We presented Transformer-Guided-Attn, a method to explicitly guide the attention heads of the Transformer using role-specific masks. The motivation of this explicit guidance is to force the heads to spread their attention on different parts of the input with the aim of reducing redundancy among the heads. Our experiments demonstrated that incorporating multiple role masks into multi-head attention can consistently improve performance on both classification and machine translation tasks.

As future work, we plan to explore additional roles for masking, as well as evaluating the impact of including it for pre-training language representation models such as BERT [5].

Acknowledgments. This work is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 721321 (QUARTZ project) and No. 893667 (METER project).

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1409.0473
2. Beltagy, I., Cohan, A., Lo, K.: Scibert: pretrained contextualized embeddings for scientific text. CoRR abs/1903.10676 (2019). http://arxiv.org/abs/1903.10676
3. Bin, Y., Yang, Y., Shen, F., Xu, X., Shen, H.T.: Bidirectional long-short term memory for video description. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 436–440 (2016)
4. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? an analysis of BERT's attention. arXiv preprint arXiv:1906.04341 (2019)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
6. Ding, Y., Liu, Y., Luan, H., Sun, M.: Visualizing and understanding neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1150–1159 (2017)
7. Hoover, B., Strobelt, H., Gehrmann, S.: exbert: a visual analysis tool to explore learned representations in transformers models. arXiv preprint arXiv:1910.05276 (2019)
8. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguist. **8**, 64–77 (2020)
9. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: a lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations (2020)
10. Lioma, C., Blanco, R.: Part of speech based term weighting for information retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 412–423. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00958-7_37

11. Lioma, C., van Rijsbergen, C.J.K.: Part of speech n-grams and information retrieval. French Review of Applied Linguistics, Special issue on Information Extraction and Linguistics XII I(2008/1), 9–22 (2008). https://www.cairn-int.info/article-E_RFLA_131_0009-part-of-speech-n-grams-and-information.htm
12. Liu, Y., et al.: Roberta: a robustly optimized Bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
13. Ma, X., et al.: A tensorized transformer for language modeling. In: Advances in Neural Information Processing Systems, pp. 2229–2239 (2019)
14. Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? In: Advances in Neural Information Processing Systems, pp. 14014–14024 (2019)
15. Sennrich, R., Haddow, B.: Linguistic input features improve neural machine translation. In: Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11–12, Berlin, Germany, pp. 83–91. The Association for Computer Linguistics (2016). https://doi.org/10.18653/v1/w16-2209
16. Sennrich, R., Haddow, B., Birch, A.: Edinburgh neural machine translation systems for WMT 16. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 371–376. Association for Computational Linguistics, Berlin, Germany, August 2016. https://www.aclweb.org/anthology/W16-2323
17. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: DISAN: directional self-attention network for RNN/CNN-free language understanding. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
18. Strubell, E., Verga, P., Andor, D., Weiss, D., McCallum, A.: Linguistically-informed self-attention for semantic role labeling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5027–5038. Association for Computational Linguistics, Brussels, Belgium, October–November 2018. https://doi.org/10.18653/v1/D18-1548, https://www.aclweb.org/anthology/D18-1548
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
20. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. pp. 5797–5808. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/p19-1580
21. Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., Simonsen, J.G.: Encoding word order in complex embeddings. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net (2020). https://openreview.net/forum?id=Hke-WTVtwr
22. Wang, D., Simonsen, J.G., Larsen, B., Lioma, C.: The Copenhagen team participation in the factuality task of the competition of automatic identification and verification of claims in political debates of the clef-2018 fact checking lab. CLEF (Working Notes) **2125** (2018)
23. Wu, W., Wang, H., Liu, T., Ma, S.: Phrase-level self-attention networks for universal sentence encoding. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3729–3738 (2018)
24. Zhang, Z., et al.: Semantics-aware BERT for language understanding. CoRR abs/1909.02209 (2019). http://arxiv.org/abs/1909.02209

6. *Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings*

Dongsheng Wang, Qiuchi Li, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. 2019. Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings. In Companion Proceedings of The 2019 World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 317–323. DOI:https://doi.org/10.1145/3308560.3316584 [170]

Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings

Dongsheng Wang
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
wang@di.ku.dk

Qiuchi Li
Department of Information Engineering
University of Padua
Padova, Italy
qiuchili@dei.unipd.it

Lucas Chaves Lima
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
lcl@di.ku.dk

Jakob Grue Simonsen
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
simonsen@di.ku.dk

Christina Lioma
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
c.lioma@di.ku.dk

ABSTRACT

When the meaning of a phrase cannot be inferred from the individual meanings of its words (e.g., *hot dog*), that phrase is said to be *non-compositional*. Automatic compositionality detection in multi-word phrases is critical in any application of semantic processing, such as search engines [9]; failing to detect non-compositional phrases can hurt system effectiveness notably. Existing research treats phrases as either compositional or non-compositional in a deterministic manner. In this paper, we operationalize the viewpoint that compositionality is contextual rather than deterministic, i.e., that whether a phrase is compositional or non-compositional depends on its context. For example, the phrase "green card" is compositional when referring to a green colored card, whereas it is non-compositional when meaning permanent residence authorization. We address the challenge of detecting this type of contextual compositionality as follows: given a multi-word phrase, we enrich the word embedding representing its semantics with evidence about its global context (terms it often collocates with) as well as its local context (narratives where that phrase is used, which we call *usage scenarios*). We further extend this representation with information extracted from external knowledge bases. The resulting representation incorporates both localized context and more general usage of the phrase and allows to detect its compositionality in a non-deterministic and contextual way. Empirical evaluation of our model on a dataset of phrase compositionality[1], manually collected by crowdsourcing contextual compositionality assessments, shows that our model outperforms state-of-the-art baselines notably on detecting phrase compositionality.

[1]https://github.com/dswang2011/ImprovedRankedList/tree/master/input

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316584

CCS CONCEPTS

• **Information systems** → *Data encoding and canonicalization.*

KEYWORDS

Compositionality detection; Knowledge base; Word embedding

ACM Reference Format:
Dongsheng Wang, Qiuchi Li, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. 2019. Contextual Compositionality Detection with External Knowledge Bases and Word Embeddings. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3316584

1 INTRODUCTION

Automatic compositionality detection refers to the automatic assessment of the extent to which the meaning of a multi-word phrase is decomposable into the meanings of its constituents words and their combination. For example, while *brown dog* is a fully compositional phrase meaning a dog of brown color, *hot dog* is a non-compositional phrase denoting a type of food. Compositionality plays a vital role in word embeddings because a non-decomposable phrase should, in principle, be treated as a single word instead of a bag of word (BOW) in word embedding approaches.

A typical line of research in automatic compositionality detection is to "perturb" the input phrase by replacing one of its constituent words at a time with its synonym, and then to measure the semantic distance between the original phrase and the perturbed phrase set [8]. The larger this distance, the less compositional the original phrase. For instance, *hot dog* would be perturbed to *warm dog* and *hot canine*. The semantic distance between the original phrase and its two perturbations is high, indicating that they denote different concepts; hence *hot dog* is non-compositional. However, the phrase *brown dog* would be perturbed to *hazel dog* and *brown canine*, which have a shorter semantic distance to *brown dog*, indicating that it is compositional.

In this paper, we posit that the compositionality of a phrase is not dichotomous or deterministic, but instead varies across scenarios. For instance, *heavy metal* could refer to a dense metal that is toxic,

which is compositional, but it could also be non-compositional when it refers to a genre of music. Previous work acknowledges this property of compositionality theoretically [8], but no operational models implementing this have been presented to this day.

Given a multi-word phrase as input, we reason that the phrase is used in some narrative, e.g., a query, sentence, snippet, document, etc. We refer to this narrative as *usage scenario* of the phrase. We combine evidence extracted from this usage scenario of the phrase with the global context (frequently co-occurring terms) of the phrase and use this to enrich the word embedding representation of the phrase. We linearly combine the weights of the tokens that are obtained from the usage scenario and the global context. We further extend this representation with information extracted from external knowledge bases.

We evaluate our model on a large dataset of phrases which are labeled as per five degrees of compositionality under various usage scenarios. We find that our model outperforms state-of-the-art baselines notably on identifying phrase compositionality. Our contributions are as follows:

- A novel model that detects phrase compositionality under different contexts and that outperforms the state of the art performance in the area.
- A benchmarking dataset of contextualized compositionality detection, that we make publicly available to the community.

2 RELATED WORK ON AUTOMATIC COMPOSITIONALITY DETECTION

Compositionality detection mainly focuses on the semantic distance or similarity calculation between a given phrase and its component words or its perturbations under a corpus or dictionary. Earlier approaches mostly estimate the similarity between the original phrase and its component words. For example, Baldwin et al. [1], and Katz and Giesbrecht [6] employ Latent Semantic Analysis (LSA) to calculate the semantic similarity (and hence to measure compositionality). Venkatapathy and Joshi [16] extended this by adding collocation features, e.g., phrase frequency, point-wise mutual information, extracted from the British National Corpus.

More recent work estimates the similarity between a phrase and perturbed versions of that phrase where the words are replaced, one at a time, by their synonyms. For instance, Kiela and Clark [7] compute the semantic distance between a phrase and its perturbation, using cosine similarity, which measures a phrase weight by pointwise-multiplication vectors of its terms. Lioma et al. [10] calculate the semantic distance with Kullback-Leibler divergence based on a language model; and, in subsequent work, Lioma et al. [8] represent the original phrase and its perturbations as ranked lists, and measure their correlation or distance.

A promising line of work uses word embeddings and deep artificial neural networks for compositionality detection. Salehi [15] employs the word-based skip-gram model for learning non-compositional phrases, treating phrases as individual tokens with vectorial composition functions. Hashimoto and Tsuruoka [5] adopt syntactic features including word index, frequency and PMI of a phrase and its components words to learn the embeddings. Yazdani et al. [17] utilize a polynomial projection function and deep artificial neural networks to learn the semantic composition and detect non-compositional phrases like those that stand out as outliers, assuming that the majority are compositional.

Closer to our work, Salehi el. [14] use Wiktionary and utilize the definition, synonyms, and translations of Wiktionary to detect non-compositional components. Specifically, they analyze the lexical overlap between the definition of a phrase and its component words to measure compositionality. They assume that multi-word phrases are included in Wiktionary, while there is no guarantee for perfect coverage of the dictionary. Unlike this approach, we use Wiktionary together with DBPedia as a structured knowledge base to represent the contextual semantics of phrases.

To our knowledge, no prior work has operationalized the compositionality of a phrase as contextual.

3 OUR CONTEXTUAL REPRESENTATION MODEL FOR COMPOSITIONALITY DETECTION

3.1 Problem Formulation

Given an input phrase p and its accompanying usage scenario s, the aim is to compute the compositionality score $Score(p)$ of phrase p with respect to usage scenario s. We follow the substitution-based line of work [7], which (a) generates perturbations of the input phrase p by substituting one word at a time with its synonym, (b) builds a semantic representation (a vector of its co-occurring terms) separately for the input phrase p and each perturbed phrase, and (c) uses the distance between the vectors of the input phrase and its perturbations to approximate the compositionality of the input phrase: the higher the distance, the less compositional the input phrase. This substitution-based line of work does not accommodate the usage scenario of the input phrase or its perturbations. The vectors of co-occurring terms are computed on one corpus, and hence these vectors represent the *global* distributional semantics of the input phrase and its perturbations. We extend this line of work by incorporating the *local* usage scenario of the phrase and its perturbations. We furthermore enrich these representations using external knowledge bases KBs. We describe this next.

Figure 1 shows how the phrase and scenario are fed into the external corpus and knowledge base in a sequential manner in the architecture, which we refer to as contextual representation model (CRM).

3.2 Building Global and Local Phrase Context

Global Phrase Context. In Natural Language Processing (NLP), the distributional semantics of an input word are computed by fixing a natural number n and, for each occurrence of a word in some corpus, finding the n words occurring immediately before, and n words occurring immediately after each occurrence of the input word (called *context window*). If there is a total of N context windows for a word, its distributional semantics in vector form can be calculated by using all these N windows. Because this is a global representation of the word's distributional semantics across the whole corpus, the vector is called a *globalized* vector. A general word embedding (e.g., word2vec) is comparable to such global context. Concretized representations of this globalized vector can be

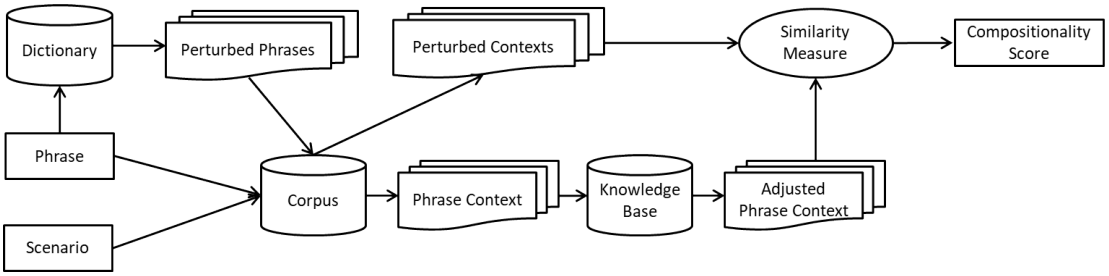


Figure 1: The diagram for the CRM sequential framework.

calculated with, e.g. ranked lists or word embeddings, as described in section 4.1.

Local Phrase Context. We aim to incorporate a representation of the local usage scenario of the input phrase (*local phrase context*) into the above described global phrase context. This representation will not be in the vector directly, because usage scenarios are typically extremely short (in terms of words), which may strongly bias the contextual representation of the phrase. Therefore, we rank all the global context windows of the phrase according to their similarity to the usage scenario of the phrase, and we select the top K most similar context windows to the usage scenario. These top K context windows are used to build the local usage scenario context representation of the phrase. Then, we linearly combine the global representation of the phrase (i.e., by taking all N context windows) and the local usage scenario representation of the phrase (i.e., the top K context windows) to acquire the *localized phrase context*. The ranking score is the similarity between the usage scenario s and a window W_i, i.e. $\text{sim}_i = \text{similarity}(W_i, s) \in [0, 1]$. The details of how the similarity score is computed are introduced in Section 4.

In the above, the value of K is determined by the length of the usage scenario as follows:

$$K = \max\left(\frac{N}{2^{\text{length}(s)}}, M\right) \tag{1}$$

where N is the total number of context windows that contain phrase p in the corpus; length(s) is the number of words in usage scenario s excluding the original phrase; and M is a threshold. We explain these next.

We posit that $2^{\text{length}(s)}$ indicates the degree of shrinking: the longer the usage scenario is, the smaller number of windows will be shrunk. The reason behind this is that the longer the usage scenario is, the more semantics it contains and subsequently fewer specific windows we are supposed to be capable of locating on. For instance, if the usage scenario is empty with length(s) = 0, then it returns the entire N windows of p with no shrinking performed; if the usage scenario has three words with $2^{length(s)} = 8$, it only collects the top $\frac{1}{8}$ of all the windows ($K = \frac{N}{8}$). Note that K depends on the usage scenario length, and is not fixed as a threshold of the similarity values. Since the similarity values may vary drastically in between $[0, 1]$ for different usage scenarios, we argue that our method is more robust to such variations. Furthermore, an empirical threshold of M is introduced to guarantee at least M windows will be selected anyhow. To this end, the localized usage scenario context of a phrase is given by:

$$C(p, s) = \alpha \frac{\sum_{i=1}^{N} \mathcal{R}(W_i)}{N} + (1 - \alpha) \frac{\sum_{j=1}^{K} \mathcal{R}(W_j)}{K} \tag{2}$$

where a is a weight parameter between 0-1 indicating the weight of global vector, and the remaining of $1 - a$ corresponds to the contribution of the localized vector; $\mathcal{R}$ denotes the semantic representation for W. In this paper, we represent a phrase as a ranked list of words and word embedding, so the same symbol $\mathcal{R}$ is adopted to denote both. The approach to calculate $\mathcal{R}$ is described in section 4.1.

3.3 Enriching Global and Local Phrase Context with Knowledge Bases

We enrich the global and local context representations of the input phrase with information extracted from external knowledge bases. We describe this next.

We reason that the corpus used to extract the global and local contexts has good coverage of various but not all possible usage scenarios of the phrase. A knowledge base is expected to contain more comprehensive, declarative information about the phrase, e.g., entities and phrase senses with categorized information. We, therefore, enrich the global and local phrase contexts extracted from the corpus with phrase information extracted from external knowledge bases.

Given an input phrase, we collect all candidate senses and entities (uniformly referred to as *candidates* in this paper) by searching the following properties (and associated values) from the knowledge bases: the properties *dbpedia:redirects*, *dbpedia:disambiguation* and their propagation relation with *dbpedia:name* and *rdfs:label*. The associated resources in these retrieved triples result in a set of candidates. Then, for each candidate, the values of *rdfs:label*, *dbpedia:abstract* and *rdf:type* are concatenated as the context for that candidate, excluding the title (which is mostly the phrase name). We also use the interface[2] to retrieve senses from Wiktionary and merge them into the same candidate set for that given input.

Most phrases only contain a limited number of candidates, and different candidates of the same phrase can have entirely different meanings or be distinct entities. We hence investigate a sequential way to incorporate the knowledge base into the phrase contextual representation, as follows. First, the phrase is fed into the knowledge base to find all candidate articles. Then, the candidates are ranked

[2]https://dkpro.github.io/dkpro-jwktl/

according to how similar they are to the localized phrase context. Those with similarity values above a certain threshold are identified as the matched candidates, denoted as $\{D_i, sim_i\}_{i=1}^{n}$, where D_i and sim_i refer to the i^{th} matched candidate with similarity value $\text{sim}_i \in [0, 1]$. A linear combination of the localized phrase context and the candidate articles is then conducted to compute the adjusted phrase context as follows:

$$C(p, D) = \lambda C(p, s) + (1 - \lambda) \sum_{i=1}^{n} w_i \mathcal{R}(D_i), \tag{3}$$

where $w_i = \frac{\text{sim}_i}{\sum_{i=1}^{n} \text{sim}_i}$ is the normalized similarity score for the i^{th} candidate article D_i while $\mathcal{R}(D_i)$ denotes the semantic representation for D_i. Since KB contains well-defined knowledge of words, we use a weighted sum of the matched candidates, instead of a simple average of matched contexts in the text corpus.

The knowledge base we employ consists of DBpedia, Wiktionary, and Wordnet. DBpedia is constructed by extracting structured information from Wikipedia. The English version of the DBpedia contains 4.58 million entries, of which 4.22 million are classified and managed under one consistent ontology. Wiktionary is a multilingual, web-based, freely available dictionary, thesaurus and phrase book, designed as the lexical companion to Wikipedia. Volunteers collaboratively construct Wiktionary, so there are no specialized qualifications necessary.

3.4 Non-linear combination

This section represents an approach of non-linear combination as a companion to the linear combination approach introduced in section 3.2 and 3.3. In addition to the weight parameter oriented design of the linear combination, we also employ a non-linear sigmoid function in RNNs (recurrent neural networks), which resolves the arrangement of the combining order for context inputs. In other words, RNNs take into consideration the feedback from the previous context vector back and forth, leading to numerous applications [3, 4]. Specifically, we train a neural network model using the Keras library to identify the compositionality label of each phrase. We encode the semantics adopting pre-trained word embeddings - word2vec [11] as word representations, a recurrent neural network with LSTM cells as the model, and cross-entropy as the loss function. As an optimizer, we utilize Adam optimizer for training the model.

In a realistic scenario (also represented in our dataset) there are fewer non-compositional than compositional phrases. This situation resembles the class imbalance issue which happens when one class (or label) is represented by most of the examples while the other one is represented just by a few. Therefore, we adopt re-sampling strategies to tackle this problem.

3.5 Compositionality Detection

Here we introduce our proposed method for compositionality detection with Algorithm 1. Given a phrase p of length l and its usage scenario s in a large corpus $Corp$, we compute its compositionality score through the following steps:

(1) Obtain localized phrase context through Eq. 1 and 2. The usage scenario of a phrase is the critical information, and the idea behind this step is to smooth the scenario context representation with the original phrase representation, as shown in line 8 in Algorithm 1.

(2) Adjust phrase context with a knowledge base. The knowledge base is fed to adjust the localized phrase context where we adopt Eq. 3 to encode the information (from line 10 to line 12).

(3) Obtain a perturbed phrase set. For each term in the phrase, we find its synonyms in WordNet. We then generate the set of perturbed phrases $S(p)$ as: S(p) = {$\hat{p}$ where $\hat{p}$ = l-1 terms of p plus a synonym of the remaining term of p}, from line 3 to line 6.

(4) Construct a perturbation representation set. For each perturbed phrase $\hat{p}$ in S(p), the corresponding representation $C(\hat{p})$ is composed of all windows of $\hat{p}$ from the corpus, and is added to the perturbation list $Q(L_{\hat{p}})$ from line 14 to line 16. Note that we do not combine context from KB for perturbations.

(5) Compute the compositionality score for the input phrase, shown in line 17, using the following equation:

$$score(p) = \frac{\sum_{\hat{p} \in S(p)} sim(C(p, D), C(\hat{p}))}{|S(p)|} \tag{4}$$

Input: Phrase p with length l
Input: Usage scenario s for p
Input: Corpus $Corp$
Input: Knowledge Base - DBPedia
Input: Similarity threshold - $thred$
Output: Compositionality score $comp(p)$
1: Set of perturbed phrase S($\hat{p}$) $\leftarrow \varnothing$
2: Find synonym $\hat{t}$ of each term $t \in p$
3: **for** each $\hat{t}$ **do**
4: Perturbed phrase $\hat{p} \leftarrow$ { $\hat{t}$, l-1 original terms t_o }
5: Update perturbed phrase set $S(\hat{p}) \leftarrow S(\hat{p}) \cup \hat{p}$
6: **end for**
7: **for** phrase $p' \in \{p \cup S(\hat{p})\}$ **do**
8: $C(p') \leftarrow$ get context terms from localized phrase context from $Corp$, smooth with Eq. 1 if it has scenario
9: **end for**
10: Find n candidate articles D_i from KB where $sim_i = similarity(s, D_i) > thred$
11: $\mathcal{R}(D_i) \leftarrow$ semantic representation of D_i
12: $C(p, D) \leftarrow$ linear combined context $\lambda C(p) + (1 - \lambda) \sum_{i=1}^{n} \frac{sim_i}{\sum_{i=1}^{n} sim_i} \mathcal{R}(D_i)$
13: $Q(L_{\hat{p}}) \leftarrow \varnothing$
14: **for** each perturbed phrase $\hat{p} \in S(\hat{p})$ **do**
15: $Q(L_{\hat{p}}) \leftarrow Q(L_{\hat{p}}) \cup C(\hat{p})$
16: **end for**
17: **return** $\frac{1}{|Q(L_{\hat{p}})|} \sum_{C(\hat{p}) \in Q(L_{\hat{p}})} Similarity(C(\hat{p}), C(p, D))$

Algorithm 1: Algorithm of contextual compositionality detection

4 IMPLEMENTATION

4.1 Semantic Representation

The semantic representation of a context (a phrase, a context window or a candidate content), i.e., $\mathcal{R}(\cdot)$ is concretized as either a ranked list or a word embedding. For the ranked list model, we calculate the TF-IDF as weight for all the tokens, rank them according to the weight, resulting in a ranked list of those tokens as the localized contextual representation. For the word embedding model, we use existing pre-trained word vectors - Glove [12], and represent the vector with the average of all tokens. The corpus we employ in our experiment is ClueWeb12-B13, a subset of some 50 million pages of ClueWeb12-Full dataset[3].

These two contextual representations lead to two different compositionality scores for the same model. We apply suffixes "-word embedding" or "-ranked list" in order to distinguish the way the contextual representation is computed, resulting in two distinct models, namely *CRM word embedding* and *CRM ranked list.*

4.2 Similarity Measure

In this study, we are faced with the problem of computing the similarity value between two context vectors. Here, we consider two types of similarity measures to achieve this purpose: *cosine similarity* and *Pearson correlation coefficient.*

One of the most commonly used similarity measures, *cosine similarity*, computes the cosine value of the angle between the two vectors of the same length. For two vectors $\vec{a} = [a_1, a_2, .., a_n]$ and $\vec{b} = [b_1, b_2, .., b_n]$, their cosine similarity cossim(a,b) is given below:

$$cossim(a, b) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{5}$$

The *Pearson correlation coefficient* computes the degree of correlation between two variables, each having a set of observed values. Suppose two variables X and Y are associated with two set of values $\{X_1, X_2, ..., X_n\}$ and $\{Y_1, Y_2, ..., Y_n\}$ respectively. The Pearson correlation coefficient r can, therefore, be computed as follows:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \tag{6}$$

where $\overline{X}$ and $\overline{Y}$ denote the average of X and Y respectively.

4.3 Perturbation

Here, we introduce the process to obtain the perturbations of a phrase p with length l. First, we get the synonyms for each word in the phrase. Then, we construct the whole perturbation set, which contains all phrases composed of $l-1$ words in p and a synonym of the remaining word. Suppose the i^{th} word has n_i synonyms, then the perturbation set contains $\sum_{i=1}^{l} n_i$ perturbed phrases. We then prune the perturbation set by filtering out the rare perturbed phrases in the text corpus. Basically, we compute the occurrence frequency of all perturbed phrases and pick the perturbed phrases with top K frequency values. In our study, we set K to be 7, which is derived from empirical observation of the data. Then, the final perturbation set contains 7 perturbed phrases in our study.

[3]https://lemurproject.org/clueweb12/

4.4 Parameter Settings

Contextual Windows setting: We set *window* = 20, which means it scans the previous 20 and subsequent 20 words of that phrase, with a sum of 40 words for each window. In Equation 1, we set $M = 10$, and the base 2 can also be parameter-free which can be changed into 2,3,4,etc., to increase the localization level.

Knowledge base threshold: As for the threshold of KB candidates, we set the threshold of similarity value between localized context and KB candidates as 0.5 to filter out those candidates with similarity less than 0.5.

Ranked list length: In line with the work [8], we set a maximum length of the ranked list as 1000, which means that we rank the tokens according to their TFIDF weight, and the tokens after the position of 1000 would be pruned.

Table 1: Results of different compositionality detection methods; *na* denotes *not applicable.*

Unsupervised Methods	ρ	α, λ
Baseline: Ranked list [8]	0.131	na
Baseline: Word Embedding	0.147	na
CRM ranked list	0.209	0.1,0.5
CRM Word Embedding	0.375	0.9, 0.1
Supervised Methods (20% Testing)	ρ	
RNN (LSTM cells)	0.176	na
RNN (LSTM cells) CRM	0.324	na

Training and testing: As we are working with imbalanced data, we use a random oversampling strategy. We split our data in a stratified fashion into 65% for training, 15% for validation, and 20% for testing. The re-sampling is be done after splitting the data into training and test, and only on the training data, i.e., none of the information in the test data is being used to create synthetic observations.

5 EXPERIMENT AND VALIDATION

In this section, we evaluate the effectiveness of the model presented in Section 3. Section 5.1 introduces the dataset and Section 5.2 presents the results achieved by our model.

5.1 Crowdsourcing data

We employ a dataset that consists of 1042 phrases that are noun-noun 2-term phrases [2]. In this dataset, each phrase was assessed four times using a binary scale (compositional or non-compositional). However, these phrases are assessed with a deterministic label, meaning that no scenario or context was given, and the degree of compositionality may not always be binary [13]. Therefore, we extend the dataset into a new version where each phrase is enriched with one or two scenarios if possible, by taking advantage of a

Table 2: Summary of dataset statistics.

No. Non-Compositional	43 (3.6%)
No. Mostly Non-Compositional	145 (12.1%)
No. Ambiguos Phrases	126 (10.5%)
No. Mostly Compositional	141 (12.0%)
No. Compositional	739 (61.8%)
Unique number of Phrases	1042
No. of context	1194
Average number of context by Phrase	1.146

crowdsourcing website - Figure Eight [4], and we use a graded level of compositionality. In Table 2 we summarize the dataset statistics.

We divided the assessment into two stages: for the first stage, the trustful assessors, with level 3 (highest in Figure Eight), are required to understand the various meanings of a phrase, and, if possible, create two scenarios for the same phrase. From these two scenarios, one should be compositional or as compositional as possible, and the other non-compositional or as non-compositional as possible. If the phrase can only be compositional or only be non-compositional, then they create one scenario for it. For the second stage, the assessors are required to assess the compositionality of phrases within different scenarios with one of the five graded labels: compositional, mostly-compositional, ambiguous to judge, mostly non-compositional, and non-compositional. Note that, for the first stage, the two scenarios of a phrase are not necessarily of two extreme polarities.

[4] https://www.figure-eight.com/

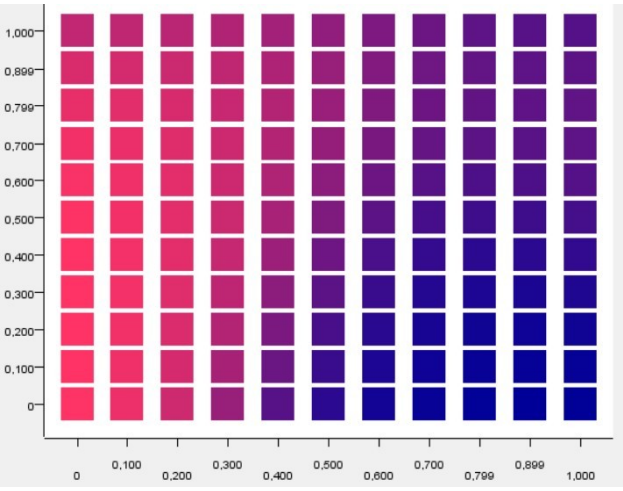


Figure 2: The grid search for Word embedding based Contextual representation. x-coordinate is α for controlling localized context and λ stands for y-coordinate, controlling the KB combining weight. Deeper blue represents higher performance whereas red indicates the opposite.

5.2 Performance and Validation

Two linear combination parameters influence the performance of our model: the combination weight α (in Eq. 2) between the vectors of a phrase and its scenario, resulting in a localized phrase context, and λ (in Eq. 3) between the localized context and knowledge base. The impacts of these two parameters on the final performance are visualized in Figure 2 and 3, corresponding to the word embedding-based and ranked list-based contextual representation respectively. α and λ denote the x and y coordinates. The colors indicate the performance, which is the correlation between the ground truth labels and the predicted labels of our models ranging from -1 to 1. The performance values are colored ranging from red to blue, representing the lowest performance to the highest.

As shown in Figure 2, the performance is negatively correlated with α while positively correlated with λ. This indicates that reducing the relative importance of localized context (right direction on x-coordinate) while enhancing the influence of knowledge base (bottom direction on y-coordinate) can improve the performance for word embedding based contextual representation. In contrast, as shown in Figure 3, if we ignore the *0* column, α is negatively correlated with the performance while λ does not have an apparent influence on the performance. This indicates that attaching higher importance to the localized context (left direction on x-coordinate) can improve the performance for the ranked list based contextual representation, while the adoption of knowledge base does not have an apparent influence to the overall performance. The first *0* column, which is shown more like an outlier, indicates that the existence of a vector of the original phrase is necessary. In other words, localized context would have relatively poor performance.

As summarized in Table 1, the performance improved from 0.147 to 0.375 for CRMs based on word embeddings (the best); from 0.131 to 0.209 for CRMs based on ranked list; and 0.176 to 0.324 for CRM based on RNN. For the word embedding based contextual representation model, relying more on the knowledge base while keeping the scenario to limited importance will lead to a high-performed model; for the ranked list based contextual representation model, on the other hand, adequately high adoption of localized context can lead to improved performance. The reason behind this can be that the knowledge base contains relatively trimmed but well-categorized information, therefore, the word embedding model can take full use of this text as informative vectors. In contrast, ranked lists, depending on tokens, work better on a large-scale corpus where they induce a large number of context windows. However, the knowledge base contains a limited number of tokens that may have little contribution to the final representation. Even though we can tune the weight of tokens from a knowledge base, it still can have limited influence in comparison to the long ranked list, which can be as long as 1000 tokens in our experiment.

For the non-linear combination where we employed the sigmoid function in RNNs, the CRM based on RNN still beats the original RNN. However, the performance is still lower than the unsupervised approaches.

6 CONCLUSION

We developed a novel method for compositionality detection where the compositionality of a phrase is contextual rather than static.

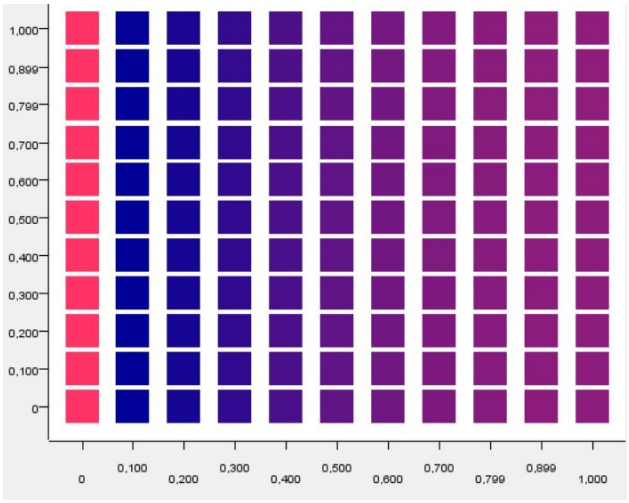


Figure 3: The grid search for ranked list based Contextual representation. x-coordinate is α for controlling localized context and y-coordinate stands for λ, controlling the KB combining weight. Deeper blue represents higher performance whereas red indicates the opposite.

Instead of considering an isolated phrase as input, we assume a phrase and its usage scenario (e.g., a query, snippet, sentence, etc.) as input, and we model a joint semantic representation of these by combining distributional semantics extracted from a corpus and additional evidence extracted from an external structured knowledge base.

Our resulting model uses word embeddings to detect compositionality, more accurately than the related state of the art. Our experiments show that for word embeddings, the usage of knowledge bases can lead to notable performance improvements.

In the future, we plan to evaluate our model on further datasets and compositionality detection scenario, e.g., Verbal Phraseological Units (VPUs).

ACKNOWLEDGEMENT

This work is supported by the Quantum Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 721321.

REFERENCES

[1] Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*. Association for Computational Linguistics, 89–96.

[2] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A Multiword Expression Dataset: Annotating Non-Compositionality and Conventionalization for English Noun Compounds. In *Proceedings of the 11th Workshop on Multiword Expressions (MWE-NAACL 2015)*. Association for Computational Linguistics.

[3] Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. [n. d.]. Neural Speed Reading with Structural-Jump-LSTM. In *Proceedings of the 2019 International Conference on Learning Representations, ICLR 2019*.

[4] Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Modelling Sequential Music Track Skips using a Multi-RNN Approach. In *Proceedings of the 2019 International Conference on Web Search and Data Mining, WSDM 2019, Sequential Skip Prediction Challenge*. in press.

[5] Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. *arXiv preprint arXiv:1603.06067* (2016).

[6] Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, 12–19.

[7] Douwe Kiela and Stephen Clark. 2013. Detecting Compositionality of Multi-Word Expressions using Nearest Neighbours in Vector Space Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1427–1432. http://aclweb.org/anthology/D13-1147

[8] Christina Lioma and Niels Dalum Hansen. 2017. A Study of Metrics of Distance and Correlation Between Ranked Lists for Compositionality Detection. *Cogn. Syst. Res.* 44, C (Aug. 2017), 40–49. https://doi.org/10.1016/j.cogsys.2017.03.001

[9] Christina Lioma, Birger Larsen, and Peter Ingwersen. 2018. To Phrase or Not to Phrase – Impact of User versus System Term Dependence Upon Retrieval. *Data and Information Management* 2, 1 (2018), 1–14.

[10] Christina Lioma, Jakob Grue Simonsen, Birger Larsen, and Niels Dalum Hansen. 2015. Non-compositional term dependence for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 595–604.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[13] Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11)*. Chiang Mai, Thailand. http://aclweb.org/anthology-new/I/I11/I11-1024.pdf

[14] Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1792–1797.

[15] Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 977–983.

[16] Sriram Venkatapathy and Aravind K Joshi. 2005. Measuring the relative compositionality of verb-noun (VN) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 899–906.

[17] Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1733–1742.

7. *MultiFC A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims*

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, & Jakob Grue Simonsen. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. 2019, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, p. 4684-4696 [15].

MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims

Isabelle Augenstein **Christina Lioma** **Dongsheng Wang** **Lucas Chaves Lima**
Casper Hansen **Christian Hansen** **Jakob Grue Simonsen**
Department of Computer Science
University of Copenhagen
{augenstein,c.lioma,wang,lcl,c.hansen,chrh,simonsen}@di.ku.dk

Abstract

We contribute the largest publicly available dataset of naturally occurring factual claims for the purpose of automatic claim verification. It is collected from 26 fact checking websites in English, paired with textual sources and rich metadata, and labelled for veracity by human expert journalists. We present an in-depth analysis of the dataset, highlighting characteristics and challenges. Further, we present results for automatic veracity prediction, both with established baselines and with a novel method for joint ranking of evidence pages and predicting veracity that outperforms all baselines. Significant performance increases are achieved by encoding evidence, and by modelling metadata. Our best-performing model achieves a Macro F1 of 49.2%, showing that this is a challenging testbed for claim veracity prediction.

1 Introduction

Misinformation and disinformation are two of the most pertinent and difficult challenges of the information age, exacerbated by the popularity of social media. In an effort to counter this, a significant amount of manual labour has been invested in fact checking claims, often collecting the results of these manual checks on fact checking portals or websites such as politifact.com or snopes.com. In a parallel development, researchers have recently started to view fact checking as a task that can be partially automated, using machine learning and NLP to automatically predict the *veracity* of claims. However, existing efforts either use small datasets consisting of naturally occurring claims (e.g. Mihalcea and Strapparava (2009); Zubiaga et al. (2016)), or datasets consisting of artificially constructed claims such as FEVER (Thorne et al., 2018). While the latter offer valuable contributions to further automatic claim verification work, they cannot replace real-world datasets.

Feature	Value
ClaimID	farg-00004
Claim	Mexico and Canada assemble cars with foreign parts and send them to the U.S. with no tax.
Label	distorts
Claim URL	https://www.factcheck.org/2018/10/factchecking-trump-on-trade/
Reason	None
Category	the-factcheck-wire
Speaker	Donald Trump
Checker	Eugene Kiely
Tags	North American Free Trade Agreement
Claim Entities	United_States, Canada, Mexico
Article Title	FactChecking Trump on Trade
Publish Date	October 3, 2018
Claim Date	Monday, October 1, 2018

Table 1: An example of a claim instance. Entities are obtained via entity linking. Article and outlink texts, evidence search snippets and pages are not shown.

Contributions. We introduce the currently largest claim verification dataset of naturally occurring claims.[1] It consists of 34,918 claims, collected from 26 fact checking websites in English; evidence pages to verify the claims; the context in which they occurred; and rich metadata (see Table 1 for an example). We perform a thorough analysis to identify characteristics of the dataset such as entities mentioned in claims. We demonstrate the utility of the dataset by training state of the art veracity prediction models, and find that evidence pages as well as metadata significantly contribute to model performance. Finally, we propose a novel model that jointly ranks evidence pages and performs veracity prediction. The best-performing model achieves a Macro F1 of 49.2%, showing that this is a non-trivial dataset with remaining challenges for future work.

[1]The dataset is found here: https://copenlu.github.io/publication/2019_emnlp_augenstein/

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4685–4697,
Hong Kong, China, November 3–7, 2019.

2 Related Work

2.1 Datasets

Over the past few years, a variety of mostly small datasets related to fact checking have been released. An overview over core datasets is given in Table 2. The datasets can be grouped into four categories (I–IV). Category I contains datasets aimed at testing how well the veracity[3] of a claim can be predicted using the claim alone, without context or evidence documents. Category II contains datasets bundled with documents related to each claim – either topically related to provide context, or serving as evidence. Those documents are, however, not annotated. Category III is for predicting veracity; they encourage retrieving evidence documents as part of their task description, but do not distribute them. Finally, category IV comprises datasets annotated for both veracity and stance. Thus, every document is annotated with a label indicating whether the document supports or denies the claim, or is unrelated to it. Additional labels can then be added to the datasets to better predict veracity, for instance by jointly training stance and veracity prediction models.

Methods not shown in the table, but related to fact checking, are stance detection for claims (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017; Augenstein et al., 2016a; Kochkina et al., 2017; Augenstein et al., 2016b; Zubiaga et al., 2018; Riedel et al., 2017), satire detection (Rubin et al., 2016), clickbait detection (Karadzhov et al., 2017), conspiracy news detection (Tacchini et al., 2017), rumour cascade detection (Vosoughi et al., 2018) and claim perspectives detection (Chen et al., 2019).

Claims are obtained from a variety of sources, including Wikipedia, Twitter, criminal reports and fact checking websites such as politifact.com and snopes.com. The same goes for documents – these are often websites obtained through Web search queries, or Wikipedia documents, tweets or Facebook posts. Most datasets contain a fairly small number of claims, and those that do not, often lack evidence documents. An exception is Thorne et al. (2018), who create a Wikipedia-based fact checking dataset. While a good testbed for developing deep neural architectures, their dataset is artificially constructed and can thus not take metadata about claims into account.

Contributions: We provide a dataset that, uniquely among extant datasets, contains a large number of *naturally occurring* claims and rich additional meta-information.

2.2 Methods

Fact checking methods partly depend on the type of dataset used. Methods only taking into account claims typically encode those with CNNs or RNNs (Wang, 2017; Pérez-Rosas et al., 2018), and potentially encode metadata (Wang, 2017) in a similar way. Methods for small datasets often use handcrafted features that are a mix of bag of word and other lexical features, e.g. LIWC, and then use those as input to a SVM or MLP (Mihalcea and Strapparava, 2009; Pérez-Rosas et al., 2018; Baly et al., 2018). Some use additional Twitter-specific features (Enayet and El-Beltagy, 2017). More involved methods taking into account evidence documents, often trained on larger datasets, consist of evidence identification and ranking following a neural model that measures the compatibility between claim and evidence (Thorne et al., 2018; Mihaylova et al., 2018; Yin and Roth, 2018).

Contributions: The latter category above is the most related to our paper as we consider evidence documents. However, existing models are not trained jointly for evidence identification, or for stance and veracity prediction, but rather employ a pipeline approach. Here, we show that a joint approach that learns to weigh evidence pages by their importance for veracity prediction can improve downstream veracity prediction performance.

3 Dataset Construction

We crawled a total of 43,837 claims with their metadata (see details in Table 11). We present the data collection in terms of selecting sources, crawling claims and associated metadata (Section 3.1); retrieving evidence pages; and linking entities in the crawled claims (Section 3.3).

3.1 Selection of sources

We crawled all active fact checking websites in English listed by Duke Reporters' Lab[4] and on the Fact Checking Wikipedia page.[5] This resulted in

[3] We use *veracity*, *claim credibility*, and *fake news* prediction interchangeably here – these terms are often conflated in the literature and meant to have the same meaning.

[4] https://reporterslab.org/fact-checking/

[5] https://en.wikipedia.org/wiki/Fact_checking

Dataset	# Claims	Labels	metadata	Claim Sources
I: Veracity prediction w/o evidence				
Wang (2017)	12,836	6	Yes	Politifact
Pérez-Rosas et al. (2018)	980	2	No	News Websites
II: Veracity				
Bachenko et al. (2008)	275	2	No	Criminal Reports
Mihalcea and Strapparava (2009)	600	2	No	Crowd Authors
Mitra and Gilbert (2015)†	1,049	5	No	Twitter
Ciampaglia et al. (2015)†	10,000	2	No	Google, Wikipedia
Popat et al. (2016)	5,013	2	Yes	Wikipedia, Snopes
Shu et al. (2018)†	23,921	2	Yes	Politifact, gossipcop.com
Datacommons Fact Check[2]	10,564	2-6	Yes	Fact Checking Websites
III: Veracity (evidence encouraged, but not provided)				
Barrn-Cedeo et al. (2018)	150	3	No	factcheck.org, Snopes
IV: Veracity + stance				
Vlachos and Riedel (2014)	106	5	Yes	Politifact, Channel 4 News
Zubiaga et al. (2016)	330	3	Yes	Twitter
Derczynski et al. (2017)	325	3	Yes	Twitter
Baly et al. (2018)	422	2	No	ara.reuters.com, verify-sy.com
Thorne et al. (2018)†	185,445	3	No	Wikipedia
V: Veracity + evidence relevancy				
MultiFC	36,534	2-40	Yes	Fact Checking Websites

Table 2: Comparison of fact checking datasets. † indicates claims are not 'naturally occuring': Mitra and Gilbert (2015) use events as claims; Ciampaglia et al. (2015) use DBPedia tiples as claims; Shu et al. (2018) use tweets as claims; and Thorne et al. (2018) rewrite sentences in Wikipedia as claims.

38 websites in total (shown in Table 11). Out of these, ten websites could not be crawled, as further detailed in Table 9. In the later experimental descriptions, we refer to the part of the dataset crawled from a specific fact checking website as a *domain*, and we refer to each website as *source*.

From each source, we crawled the ID, claim, label, URL, reason for label, categories, person making the claim (speaker), person fact checking the claim (checker), tags, article title, publication date, claim date, as well as the full text that appears when the claim is clicked. Lastly, the above full text contains hyperlinks, so we further crawled the full text that appears when each of those hyperlinks are clicked (outlinks).

There were a number of crawling issues, e.g. security protection of websites with SSL/TLS protocols, time out, URLs that pointed to pdf files instead of HTML content, or unresolvable encoding. In all of these cases, the content could not be retrieved. For some websites, no veracity labels were available, in which case, they were not selected as domains for training a veracity prediction model. Moreover, not all types of metadata (category, speaker, checker, tags, claim date, publish date) were available for all websites; and availability of articles and full texts differs as well.

We performed semi-automatic cleansing of the dataset as follows. First, we double-checked that the veracity labels would not appear in claims. For some domains, the first or last sentence of the claim would sometimes contain the veracity label, in which case we would discard either the full sentence or part of the sentence. Next, we checked the dataset for duplicate claims. We found 202 such instances, 69 of them with different labels. Upon manual inspection, this was mainly due to them appearing on different websites, with labels not differing much in practice (e.g. 'Not true', vs. 'Mostly False'). We made sure that all such duplicate claims would be in the training split of the dataset, so that the models would not have an unfair advantage. Finally, we performed some minor manual merging of label types for the same domain where it was clear that they were supposed to denote the same level of veracity (e.g. 'distorts', 'distorts the facts').

This resulted in a total of 36,534 claims with their metadata. For the purposes of fact verification, we discarded instances with labels that occur fewer than 5 times, resulting in 34,918 claims. The number of instances, as well as labels per domain, are shown in Table 6 and label names in Table 10 in the appendix. The dataset is split into a training part (80%) and a development and testing part (10% each) in a label-stratified manner. Note that

the domains vary in the number of labels, ranging from 2 to 27. Labels include both straight-forward ratings of veracity ('correct', 'incorrect'), but also labels that would be more difficult to map onto a veracity scale (e.g. 'grass roots movement!', 'misattributed', 'not the whole story'). We therefore do not postprocess label types across domains to map them onto the same scale, and rather treat them as is. In the methodology section (Section 4), we show how a model can be trained on this dataset regardless by framing this multi-domain veracity prediction task as a multi-task learning (MTL) one.

3.2 Retrieving Evidence Pages

The text of each claim is submitted verbatim as a query to the Google Search API (without quotes). The 10 most highly ranked search results are retrieved, for each of which we save the title; Google search rank; URL; time stamp of last update; search snippet; as well as the full Web page. We acknowledge that search results change over time, which might have an effect on veracity prediction. However, studying such temporal effects is outside the scope of this paper. Similar to Web crawling claims, as described in Section 3.1, the corresponding Web pages can in some cases not be retrieved, in which case fewer than 10 evidence pages are available. The resulting evidence pages are from a wide variety of URL domains, though with a predictable skew towards popular websites, such as Wikipedia or The Guardian (see Table 3 for detailed statistics).

3.3 Entity Detection and Linking

To better understand what claims are about, we conduct entity linking for all claims. Specifically, mentions of people, places, organisations, and other named entities within a claim are recognised and linked to their respective Wikipedia pages, if available. Where there are different entities with the same name, they are disambiguated. For this, we apply the state-of-the-art neural entity linking model by Kolitsas et al. (2018). This results in a total of 25,763 entities detected and linked to Wikipedia, with a total of 15,351 claims involved, meaning that 42% of all claims contain entities that can be linked to Wikipedia. Later on, we use entities as additional metadata (see Section 4.3). The distribution of claim numbers according to the number of entities they contain is shown in Figure 1. We observe that the majority of claims have one to four entities, and the maximum number of 35 entities occurs in one claim only. Out of the 25,763 entities, 2,767 are unique entities. The top 30 most frequent entities are listed in Table 4. This clearly shows that most of the claims involve entities related to the United States, which is to be expected, as most of the fact checking websites are US-based.

Domain	%
https://en.wikipedia.org/	4.425
https://www.snopes.com/	3.992
https://www.washingtonpost.com/	3.025
https://www.nytimes.com/	2.478
https://www.theguardian.com/	1.807
https://www.youtube.com/	1.712
https://www.dailymail.co.uk/	1.558
https://www.usatoday.com/	1.279
https://www.politico.com/	1.241
http://www.politifact.com/	1.231
https://www.pinterest.com/	1.169
https://www.factcheck.org/	1.09
https://www.gossipcop.com/	1.073
https://www.cnn.com/	1.065
https://www.npr.org/	0.957
https://www.forbes.com/	0.911
https://www.vox.com/	0.89
https://www.theatlantic.com/	0.88
https://twitter.com/	0.767
https://www.hoax-slayer.net/	0.655
http://time.com/	0.554
https://www.bbc.com/	0.551
https://www.nbcnews.com/	0.515
https://www.cnbc.com/	0.514
https://www.cbsnews.com/	0.503
https://www.facebook.com/	0.5
https://www.newyorker.com/	0.495
https://www.foxnews.com/	0.468
https://people.com/	0.439
http://www.cnn.com/	0.419

Table 3: The top 30 most frequently occurring URL domains.

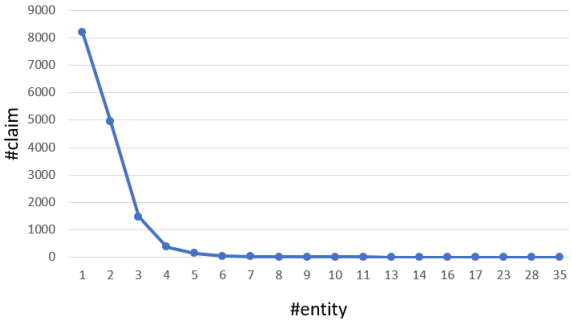


Figure 1: Distribution of entities in claims.

4 Claim Veracity Prediction

We train several models to predict the veracity of claims. Those fall into two categories: those that

Entity	Frequency
United_States	2810
Barack_Obama	1598
Republican_Party_(United_States)	783
Texas	665
Democratic_Party_(United_States)	560
Donald_Trump	556
Wisconsin	471
United_States_Congress	354
Hillary_Rodham_Clinton	306
Bill_Clinton	292
California	285
Russia	275
Ohio	239
China	229
George_W._Bush	208
Medicare_(United_States)	206
Australia	186
Iran	183
Brad_Pitt	180
Islam	178
Iraq	176
Canada	174
White_House	166
New_York_City	164
Washington,_D.C.	164
Jennifer_Aniston	163
Mexico	158
Ted_Cruz	152
Federal_Bureau_of_Investigation	146
Syria	130

Table 4: Top 30 most frequent entities listed by their Wikipedia URL with prefix omitted

only consider the claims themselves, and those that encode evidence pages as well. In addition, claim metadata (speaker, checker, linked entities) is optionally encoded for both categories of models, and ablation studies with and without that metadata are shown. We first describe the base model used in Section 4.1, followed by introducing our novel evidence ranking and veracity prediction model in Section 4.2, and lastly the metadata encoding model in Section 4.3.

4.1 Multi-Domain Claim Veracity Prediction with Disparate Label Spaces

Since not all fact checking websites use the same claim labels (see Table 6, and Table 10 in the appendix), training a claim veracity prediction model is not entirely straight-forward. One option would be to manually map those labels onto one another. However, since the sheer number of labels is rather large (165), and it is not always clear from the guidelines on fact checking websites how they can be mapped onto one another, we opt to learn how these labels relate to one another as part of the veracity prediction model. To do so, we employ the multi-task learning (MTL) approach inspired by collaborative filtering presented in Augenstein et al. (2018) (*MTL with LEL*–multitask learning with label embedding layer) that excels on pairwise sequence classification tasks with disparate label spaces. More concretely, each domain is modelled as its own task in a MTL architecture, and labels are projected into a fixed-length label embedding space. Predictions are then made by taking the dot product between the claim-evidence embeddings and the label embeddings. By doing so, the model implicitly learns how semantically close the labels are to one another, and can benefit from this knowledge when making predictions for individual tasks, which on their own might only have a small number of instances. When making predictions for individual domains/tasks, both at training and at test time, as well as when calculating the loss, a mask is applied such that the valid and invalid labels for that task are restricted to the set of known task labels.

Note that the setting here slightly differs from Augenstein et al. (2018). There, tasks are less strongly related to one another; for example, they consider stance detection, aspect-based sentiment analysis and natural language inference. Here, we have different domains, as opposed to conceptually different tasks, but use their framework, as we have the same underlying problem of disparate label spaces. A more formal problem definition follows next, as our evidence ranking and veracity prediction model in Section 4.2 then builds on it.

4.1.1 Problem Definition

We frame our problem as a multi-task learning one, where access to labelled datasets for T tasks $\mathcal{T}_1, \ldots, \mathcal{T}_T$ is given at training time with a target task $\mathcal{T}_T$ that is of particular interest. The training dataset for task $\mathcal{T}_i$ consists of N examples $X_{\mathcal{T}_i} = \{x_1^{\mathcal{T}_i}, \ldots, x_N^{\mathcal{T}_i}\}$ and their labels $Y_{\mathcal{T}_i} = \{\mathbf{y}_1^{\mathcal{T}_i}, \ldots, \mathbf{y}_N^{\mathcal{T}_i}\}$. The base model is a classic deep neural network MTL model (Caruana, 1993) that shares its parameters across tasks and has task-specific softmax output layers that output a probability distribution $\mathbf{p}^{\mathcal{T}_i}$ for task $\mathcal{T}_i$:

$$\mathbf{p}^{\mathcal{T}_i} = \text{softmax}(\mathbf{W}^{\mathcal{T}_i}\mathbf{h} + \mathbf{b}^{\mathcal{T}_i}) \quad (1)$$

where $\text{softmax}(\mathbf{x}) = e^{\mathbf{x}} / \sum_{i=1}^{\|\mathbf{x}\|} e^{\mathbf{x}_i}$, $\mathbf{W}^{\mathcal{T}_i} \in \mathbb{R}^{L_i \times h}$, $\mathbf{b}^{\mathcal{T}_i} \in \mathbb{R}^{L_i}$ is the weight matrix and bias term of the output layer of task $\mathcal{T}_i$ respectively, $\mathbf{h} \in \mathbb{R}^h$ is the jointly learned hidden rep-

resentation, L_i is the number of labels for task $\mathcal{T}_i$, and h is the dimensionality of $\mathbf{h}$. The MTL model is trained to minimise the sum of individual task losses $\mathcal{L}_1 + \ldots + \mathcal{L}_T$ using a negative log-likelihood objective.

Label Embedding Layer. To learn the relationships between labels, a Label Embedding Layer (LEL) embeds labels of all tasks in a joint Euclidian space. Instead of training separate softmax output layers as above, a label compatibility function $c(\cdot, \cdot)$ measures how similar a label with embedding $\mathbf{l}$ is to the hidden representation $\mathbf{h}$:

$$c(\mathbf{l}, \mathbf{h}) = \mathbf{l} \cdot \mathbf{h} \tag{2}$$

where $\cdot$ is the dot product. Padding is applied such that l and h have the same dimensionality. Matrix multiplication and softmax are used for making predictions:

$$\mathbf{p} = \text{softmax}(\mathbf{Lh}) \tag{3}$$

where $\mathbf{L} \in \mathbb{R}^{(\sum_i L_i) \times l}$ is the label embedding matrix for all tasks and l is the dimensionality of the label embeddings. We apply a task-specific mask to $\mathbf{L}$ in order to obtain a task-specific probability distribution $\mathbf{p}^{\mathcal{T}_i}$. The LEL is shared across all tasks, which allows the model to learn the relationships between labels in the joint embedding space.

4.2 Joint Evidence Ranking and Claim Veracity Prediction

So far, we have ignored the issue of how to obtain claim representation, as the base model described in the previous section is agnostic to how instances are encoded. A very simple approach, which we report as a baseline, is to encode claim texts only. Such a model ignores evidence for and against a claim, and ends up guessing the veracity based on surface patterns observed in the claim texts.

We next introduce two variants of evidence-based veracity prediction models that encode 10 pieces of evidence in addition to the claim. Here, we opt to encode search snippets as opposed to whole retrieved pages. While the latter would also be possible, it comes with a number of additional challenges, such as encoding large documents, parsing tables or PDF files, and encoding images or videos on these pages, which we leave to future work. Search snippets also have the benefit that they already contain summaries of the part of the page content that is most related to the claim.

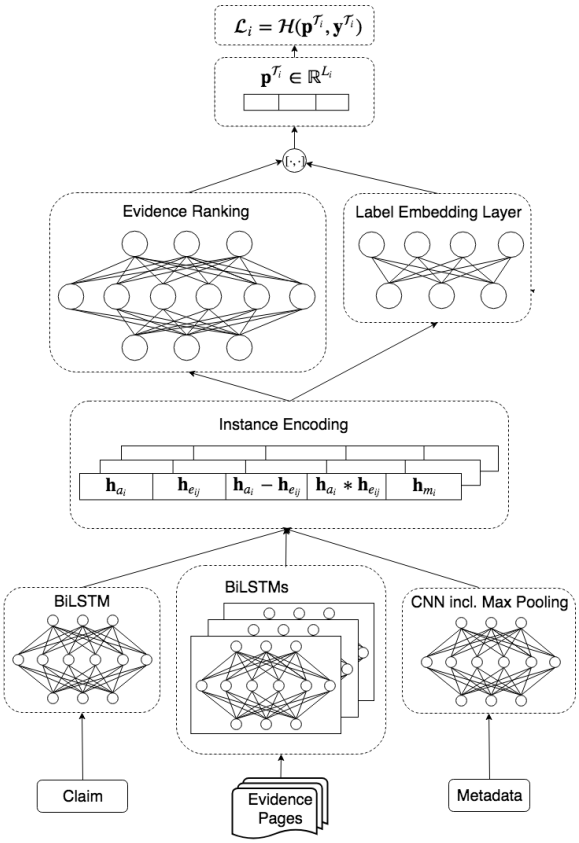


Figure 2: The Joint Veracity Prediction and Evidence Ranking model, shown for one task.

4.2.1 Problem Definition

Our problem is to obtain encodings for N examples $X_{\mathcal{T}_i} = \{x_1^{\mathcal{T}_i}, \ldots, x_N^{\mathcal{T}_i}\}$. For simplicity, we will henceforth drop the task superscript and refer to instances as $X = \{x_1, \ldots, x_N\}$, as instance encodings are learned in a task-agnostic fashion. Each example further consists of a claim a_i and $k = 10$ evidence pages $E_k = \{e_{1_0}, \ldots, e_{N_{10}}\}$.

Each claim and evidence page is encoded with a BiLSTM to obtain a sentence embedding, which is the concatenation of the last state of the forward and backward reading of the sentence, i.e. $\mathbf{h} = BiLSTM(\cdot)$, where $\mathbf{h}$ is the sentence embedding.

Next, we want to combine claims and evidence sentence embeddings into joint instance representations. In the simplest case, referred to as model variant *crawled_avg*, we mean average the BiLSTM sentence embeddings of all evidence pages (signified by the overline) and concatenate those with the claim embeddings, i.e.

$$\mathbf{s}_{g_i} = [\mathbf{h}_{a_i}; \overline{\mathbf{h}_{E_i}}] \tag{4}$$

where s_{g_i} is the resulting encoding for training example i and $[\cdot;\cdot]$ denotes vector concatenation.

However, this has the disadvantage that all evidence pages are considered equal.

Evidence Ranking The here proposed alternative instance encoding model, *crawled_ranked*, which achieves the highest overall performance as discussed in Section 5, learns the compatibility between an instance's claim and each evidence page. It ranks evidence pages by their utility for the veracity prediction task, and then uses the resulting ranking to obtain a weighted combination of all claim-evidence pairs. No direct labels are available to learn the ranking of individual documents, only for the veracity of the associated claim, so the model has to learn evidence ranks implicitly.

To combine claim and evidence representations, we use the matching model proposed for the task of natural language inference by Mou et al. (2016) and adapt it to combine an instance's claim representation with each evidence representation, i.e.

$$s_{r_{i_j}} = [\mathbf{h}_{a_i}; \mathbf{h}_{e_{i_j}}; \mathbf{h}_{a_i} - \mathbf{h}_{e_{i_j}}; \mathbf{h}_{a_i} \cdot \mathbf{h}_{e_{i_j}}] \quad (5)$$

where $s_{r_{i_j}}$ is the resulting encoding for training example i and evidence page j , $[\cdot;\cdot]$ denotes vector concatenation, and $\cdot$ denotes the dot product.

All joint claim-evidence representations $\mathbf{s}_{r_{i_0}}, \ldots, \mathbf{s}_{r_{i_{10}}}$ are then projected into the binary space via a fully connected layer FC, followed by a non-linear activation function f, to obtain a soft ranking of claim-evidence pairs, in practice a 10-dimensional vector,

$$\mathbf{o}_i = [f(\mathrm{FC}(s_{r_{i_0}})); \ldots; f(\mathrm{FC}(s_{r_{i_{10}}}))] \quad (6)$$

where $[\cdot;\cdot]$ denotes concatenation.

Scores for all labels are obtained as per (6) above, with the same input instance embeddings as for the evidence ranker, i.e. $s_{r_{i_j}}$. Final predictions for all claim-evidence pairs are then obtained by taking the dot product between the label scores and binary evidence ranking scores, i.e.

$$\mathbf{p}_i = \mathrm{softmax}(c(\mathbf{l}, \mathbf{s}_{\mathbf{r_i}}) \cdot \mathbf{o}_i) \quad (7)$$

Note that the novelty here is that, unlike for the model described in Mou et al. (2016), we have no direct labels for learning weights for this matching model. Rather, our model has to implicitly learn these weights for each claim-evidence pair in an end-to-end fashion given the veracity labels.

Model	Micro F1	Macro F1
claim-only	0.469	0.253
claim-only_embavg	0.384	0.302
crawled-docavg	0.438	0.248
crawled_ranked	0.613	0.441
claim-only + meta	0.494	0.324
claim-only_embavg + meta	0.418	0.333
crawled-docavg + meta	0.483	0.286
crawled_ranked + meta	**0.625**	**0.492**

Table 5: Results with different model variants on the test set, 'meta' means all metadata is used.

4.3 Metadata

We experiment with how useful claim metadata is, and encode the following as one-hot vectors: speaker, category, tags and linked entities. We do not encode 'Reason' as it gives away the label, and do not include 'Checker' as there are too many unique checkers for this information to be relevant. The claim publication date is potentially relevant, but it does not make sense to merely model this as a one-hot feature, so we leave incorporating temporal information to future work. Since all metadata consists of individual words and phrases, a sequence encoder is not necessary, and we opt for a CNN followed by a max pooling operation as used in Wang (2017) to encode metadata for fact checking. The max-pooled metadata representations, denoted h_m, are then concatenated with the instance representations, e.g. for the most elaborate model, *crawled_ranked*, these would be concatenated with $s_{cr_{i_j}}$.

5 Experiments

5.1 Experimental Setup

The base sentence embedding model is a BiLSTM over all words in the respective sequences with randomly initialised word embeddings, following Augenstein et al. (2018). We opt for this strong baseline sentence encoding model, as opposed to engineering sentence embeddings that work particularly well for this dataset, to showcase the dataset. We would expect pre-trained contextual encoding models, e.g. ELMO (Peters et al., 2018), ULMFit (Howard and Ruder, 2018), BERT (Devlin et al., 2018), to offer complementary performance gains, as has been shown for a few recent papers (Wang et al., 2018; Rajpurkar et al., 2018).

For claim veracity prediction without evidence documents with the MTL with LEL model, we use the following sentence encoding variants: *claim-*

only, which uses a BiLSTM-based sentence embedding as input, and *claim-only_embavg*, which uses a sentence embedding based on mean averaged word embeddings as input.

We train one multi-task model per task (i.e., one model per domain). We perform a grid search over the following hyperparameters, tuned on the respective dev set, and evaluate on the correspoding test set (final settings are underlined): word embedding size [64, 128, 256], BiLSTM hidden layer size [64, 128, 256], number of BiLSTM hidden layers [1, 2, 3], BiLSTM dropout on input and output layers [0.0, 0.1, 0.2, 0.5], word-by-word-attention for BiLSTM with window size 10 (Bahdanau et al., 2014) [True, False], skip-connections for the BiLSTM [True, False], batch size [32, 64, 128], label embedding size [16, 32, 64]. We use ReLU as an activation function for both the BiLSTM and the CNN. For the CNN, the following hyperparameters are used: number filters [32], kernel size [32]. We train using cross-entropy loss and the RMSProp optimiser with initial learning rate of 0.001 and perform early stopping on the dev set with a patience of 3.

5.2 Results

For each domain, we compute the Micro as well as Macro F1, then mean average results over all domains. Core results with all vs. no metadata are shown in Table 5. We first experiment with different base model variants and find that label embeddings improve results, and that the best proposed models utilising multiple domains outperform single-task models (see Table 8). This corroborates the findings of Augenstein et al. (2018). Per-domain results with the best model are shown in Table 6. Domain names are from hereon after abbreviated for brevity, see Table 11 in the appendix for correspondences to full website names. Unsurprisingly, it is hard to achieve a high Macro F1 for domains with many labels, e.g. tron and snes. Further, some domains, surprisingly mostly with small numbers of instances, seem to be very easy – a perfect Micro and Macro F1 score of 1.0 is achieved on ranz, bove, buca, fani and thal. We find that for those domains, the verdict is often already revealed as part of the claim using explicit wording.

Claim-Only vs. Evidence-Based Veracity Prediction. Our evidence-based claim veracity prediction models outperform claim-only veracity prediction models by a large margin. Unsurprisingly, *claim-only_embavg* is outperformed by *claim-only*. Further, *crawled_ranked* is our best-performing model in terms of Micro F1 and Macro F1, meaning that our model captures that not every piece of evidence is equally important, and can

Domain	# Insts	# Labs	Micro F1	Macro F1
ranz	21	2	1.000	1.000
bove	295	2	1.000	1.000
abbc	436	3	0.463	0.453
huca	34	3	1.000	1.000
mpws	47	3	0.667	0.583
peck	65	3	0.667	0.472
faan	111	3	0.682	0.679
clck	38	3	0.833	0.619
fani	20	3	1.000	1.000
chct	355	4	0.550	0.513
obry	59	4	0.417	0.268
vees	504	4	0.721	0.425
faly	111	5	0.278	0.5
goop	2943	6	0.822	0.387
pose	1361	6	0.438	0.328
thet	79	6	0.55	0.37
thal	163	7	1.000	1.000
afck	433	7	0.357	0.259
hoer	1310	7	0.694	0.549
para	222	7	0.375	0.311
wast	201	7	0.344	0.214
vogo	654	8	0.594	0.297
pomt	15390	9	0.321	0.276
snes	6455	12	0.551	0.097
farg	485	11	0.500	0.140
tron	3423	27	0.429	0.046
avg		7.17	0.625	0.492

Table 6: Total number of instances and unique labels per domain, as well as per-domain results with model *crawled_ranked* + *meta*, sorted by label size

Metadata	Micro F1	Macro F1
None	**0.627**	0.441
Speaker	0.602	0.435
+ Tags	0.608	0.460
Tags	0.585	0.461
Entity	0.569	0.427
+ Speaker	0.607	0.477
+ Tags	0.625	**0.492**

Table 7: Ablation results with base model *crawled_ranked* for different types of metadata

Model	Micro F1	Macro F1
STL	0.527	0.388
MTL	0.556	0.448
MTL + LEL	**0.625**	**0.492**

Table 8: Ablation results with *crawled_ranked* + *meta* encoding for STL vs. MTL vs. MTL + LEL training

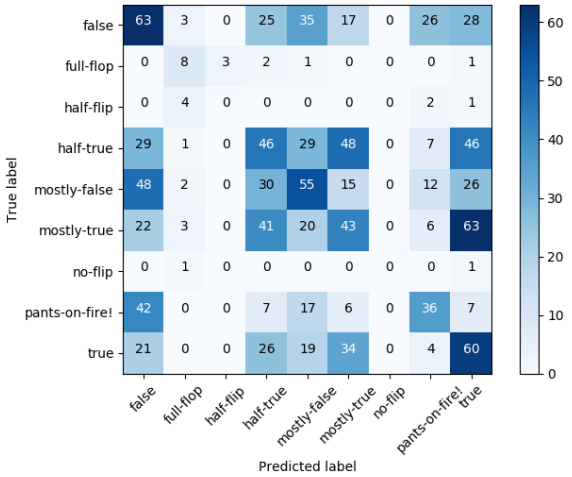


Figure 3: Confusion matrix of predicted labels with best-performing model, *crawled_ranked + meta*, on the 'pomt' domain

utilise this for veracity prediction.

Metadata. We perform an ablation analysis of how metadata impacts results, shown in Table 7. Out of the different types of metadata, topic tags on their own contribute the most. This is likely because they offer highly complementary information to the claim text of evidence pages. Only using all metadata together achieves a higher Macro F1 at similar Micro F1 than using no metadata at all. To further investigate this, we split the test set into those instances for which no metadata is available vs. those for which metadata is available. We find that encoding metadata within the model hurts performance for domains where no metadata is available, but improves performance where it is. In practice, an ensemble of both types of models would be sensible, as well as exploring more involved methods of encoding metadata.

6 Analysis and Discussion

An analysis of labels frequently confused with one another, for the largest domain 'pomt' and best-performing model *crawled_ranked + meta* is shown in Figure 3. The diagonal represents when gold and predicted labels match, and the numbers signify the number of test instances. One can observe that the model struggles more to detect claims with labels 'true' than those with label 'false'. Generally, many confusions occur over close labels, e.g. 'half-true' vs. 'mostly true'.

We further analyse what properties instances that are predicted correctly vs. incorrectly have, using the model *crawled_ranked meta*. We find that, unsurprisingly, longer claims are harder to classify correctly, and that claims with a high direct token overlap with evidence pages lead to a high evidence ranking. When it comes to frequently occurring tags and entities, very general tags such as 'government-and-politics' or 'tax' that do not give away much, frequently co-occur with incorrect predictions, whereas more specific tags such as 'brisbane-4000' or 'hong-kong' tend to co-occur with correct predictions. Similar trends are observed for bigrams. This means that the model has an easy time succeeding for instances where the claims are short, where specific topics tend to co-occur with certain veracities, and where evidence documents are highly informative. Instances with longer, more complex claims where evidence is ambiguous remain challenging.

7 Conclusions

We present a new, real-world fact checking dataset, currently the largest of its kind. It consists of 34,918 claims collected from 26 fact checking websites, rich metadata and 10 retrieved evidence pages per claim. We find that encoding the metadata as well evidence pages helps, and introduce a new joint model for ranking evidence pages and predicting veracity.

Acknowledgments

This research is partially supported by QUARTZ (721321, EU H2020 MSCA-ITN) and DABAI (5153-00004A, Innovation Fund Denmark).

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016a. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces. In *NAACL-HLT*, pages 1896–1906. Association for Computational Linguistics.

Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 2016b. USFD at SemEval-2016 Task 6: Any-Target Stance Detection on Twitter with Autoencoders. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 389–393, San Diego, California. Association for Computational Linguistics.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. 2008. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 41–48. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*.

Ramy Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 21–27. Association for Computational Linguistics.

Alberto Barrn-Cedeo, Tamer Elsayed, Reem Suwaileh, Llus Mrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. Task 2: Factuality. In *CLEF (Working Notes)*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rich Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proceedings of ICML*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *Proceedings of NAACL*.

G L Ciampaglia, P Shiralkar, L M Rocha, J Bollen, F Menczer, and A Flammini. 2015. Computational Fact Checking from Knowledge Networks. *PLoS One*, 10(6).

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Omar Enayet and Samhaa R. El-Beltagy. 2017. NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *HLT-NAACL*, pages 1163–1168. The Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *ACL (1)*, pages 328–339. Association for Computational Linguistics.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We Built a Fake News / Click Bait Filter: What Happened Next Will Blow Your Mind! In *RANLP 2017*, pages 334–343.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James R. Glass. 2018. Fact Checking in Community Forums. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press.

Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural Language Inference by Tree-Based Convolution and Heuristic Matching. In *ACL (2)*. The Association for Computer Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*, pages 2227–2237. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. The Fake News Challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. http://www.fakenewschallenge.org/. Accessed: 2019-02-14.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *CIKM*, pages 2173–2178.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *ACL (2)*, pages 784–789. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR*, abs/1707.03264.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17. Association for Computational Linguistics.

K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *ArXiv e-prints*.

Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some Like it Hoax: Automated Fake News Detection in Social Networks. In *Proceedings of the Second Workshop on Data Science for Social Good (SoGood)*, volume 1960 of *CEUR Workshop Proceedings*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *BlackboxNLP@EMNLP*, pages 353–355. Association for Computational Linguistics.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426. Association for Computational Linguistics.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Informatino Processing & Management*, 54(2):273–290.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.

Websites (Sources)	Reason
Mediabiasfactcheck	Website that checks other news websites
CBC	No pattern to crawl
apnews.com/APFactCheck	No categorical label and no structured claim
weeklystandard.com/tag/fact-check	Mostly no label, and they are placed anywhere
ballotpedia.org	No categorical label and no structured claim
channel3000.com/news/politics/reality-check	No categorical label, lack of structure, and no clear claim
npr.org/sections/politics-fact-check	No label and no clear claim (only some titles are claims)
dailycaller.com/buzz/check-your-fact	Is a subset of checkyourfact which has already been crawled
sacbee.com[6]	Contains very few labelled articles, and without clear claims
TheGuardian	Only a few websites have a pattern for labels.

Table 9: The list of websites that we did not crawl and reasons for not crawling them.

Domain	# Insts	# Labels	Labels
abbc	436	3	in-between, in-the-red, in-the-green
afck	433	7	correct, incorrect, mostly-correct, unproven, misleading, understated, exaggerated
bove	295	2	none, rating: false
chct	355	4	verdict: true, verdict: false, verdict: unsubstantiated, none
clck	38	3	incorrect, unsupported, misleading
faan	111	3	factscan score: false, factscan score: true, factscan score: misleading
faly	71	5	true, none, partly true, unverified, false
fani	20	3	conclusion: accurate, conclusion: false, conclusion: unclear
farg	485	11	false, none, distorts the facts, misleading, spins the facts, no evidence, not the whole story, unsupported, cherry picks, exaggerates, out of context
goop	2943	6	0, 1, 2, 3, 4, 10
hoer	1310	7	facebook scams, true messages, bogus warning, statirical reports, fake news, unsubstantiated messages, misleading recommendations
huca	34	3	a lot of baloney, a little baloney, some baloney
mpws	47	3	accurate, false, misleading
obry	59	4	mostly_true, verified, unobservable, mostly_false
para	222	7	mostly false, mostly true, half-true, false, true, pants on fire!, half flip
peck	65	3	false, true, partially true
pomt	15390	9	half-true, false, mostly true, mostly false, true, pants on fire!, full flop, half flip, no flip
pose	1361	6	promise kept, promise broken, compromise, in the works, not yet rated, stalled
ranz	21	2	fact, fiction
snes	6455	12	false, true, mixture, unproven, mostly false, mostly true, miscaptioned, legend, outdated, misattributed, scam, correct attribution
thet	79	6	none, mostly false, mostly true, half true, false, true
thal	74	2	none, we rate this claim false
tron	3423	27	fiction!, truth!, unproven!, truth! & fiction!, mostly fiction!, none, disputed!, truth! & misleading!, authorship confirmed!, mostly truth!, incorrect attribution!, scam!, investigation pending!, confirmed authorship!, commentary!, previously truth! now resolved!, outdated!, truth! & outdated!, virus!, fiction! & satire!, truth! & unproven!, misleading!, grass roots movement!, opinion!, correct attribution!, truth! & disputed!, inaccurate attribution!
vees	504	4	none, fake, misleading, false
vogo	653	8	none, determination: false, determination: true, determination: mostly true, determination: misleading, determination: barely true, determination: huckster propaganda, determination: false, determination: a stretch
wast	201	7	4 pinnochios, 3 pinnochios, 2 pinnochios, false, not the whole story, needs context, none

Table 10: Number of instances, and labels per domain sorted by number of occurrences

Website	Domain	Claims	Labels	Category	Speaker	Checker	Tags	Article	Claim date	Publish date	Full text	Outlinks
abc	abbc	436	436	436	-	-	436	436	-	436	436	7676
africacheck	afck	436	436	-	-	-	-	436	-	436	436	2325
altnews	-	496	-	-	-	496	-	496	-	496	496	6389
boomlive	-	302	302	-	-	-	-	302	-	302	302	6054
checkyourfact	chht	358	358	-	-	358	-	-	-	358	358	5271
climatefeedback	clck	45	45	-	-	-	-	45	-	45	45	489
crikey	-	18	18	18	-	18	18	18	-	18	18	212
factcheckni	-	36	36	36	-	-	-	36	-	-	36	151
factcheckorg	farg	512	512	512	512	512	512	512	512	512	512	8282
factly	-	77	77	-	-	-	-	77	-	-	77	658
factscan	-	115	115	-	115	-	-	-	115	115	115	1138
fullfact	-	336	336	336	-	336	-	336	-	336	336	3838
gossipcop	goop	2947	2947	-	-	2947	-	2947	-	2947	2947	12583
hoaxslayer	hoer	1310	1310	-	-	1310	-	1310	-	1310	1310	14499
huffingtonpostca	huca	38	38	-	38	38	-	38	38	38	38	78
leadstories	-	1547	1547	-	-	1547	-	1547	-	1547	1547	12015
mprnews	mpws	49	49	-	-	49	-	49	-	49	49	319
nytimes	-	17	17	-	-	17	-	17	-	17	17	271
observatory	obry	60	60	-	-	60	-	60	-	60	60	592
pandora	para	225	225	225	225	225	-	225	-	225	225	114
pesacheck	peck	67	67	-	-	67	-	67	-	67	67	521
politico	-	102	102	-	-	102	-	102	-	102	102	150
politifact_promise	pose	1361	1361	1361	1361	-	-	1361	-	1361	1361	6279
politifact_stmt	pomt	15390	15390	-	15390	-	-	-	15390	15390	15390	78543
politifact_story	-	5460	-	-	-	5460	-	-	-	5460	5460	24836
radionz	ranz	32	32	32	32	-	-	32	32	32	32	44
snopes	snes	6457	6457	6457	-	6457	-	6457	-	6457	6457	46735
swissinfo	-	20	20	20	20	20	-	20	-	20	20	40
theconversation	-	62	62	62	62	62	62	62	-	62	62	723
theferret	thet	81	81	81	81)	-	-	81	-	81(81)	81	885
theguardian	-	155	155	155	-	155	-	155	-	155	155	2600
thejournal	thal	179	179	-	-	-	-	179	-	179	179	2375
truthorfiction	tron	3674	3674	3674	-	-	3674	3674	-	3674	3674	8268
verafiles	vees	509	509	-	-	-	509	509	-	509	509	23
voiceofsandiego	vogo	660	660	-	-	-	-	660	-	660	660	2352
washingtonpost	wast	227	227	-	227	227	-	227	-	227	227	2470
wral	-	20	20	-	-	20	20	20	-	20	20	355
zimfact	-	21	21	21	21	21	-	21	-	21	21	179
Total		43837	43837	43837	43837	43837	43837	43837	43837	43837	43837	260330

Table 11: Summary statistics for claim collection. 'Domain' indicates the domain name used for the veracity prediction experiments, '–' indicates that the website was not used due to missing or insufficient claim labels, see Section 3.2.

8. *Automatic Fake News Detection: Are Models Learning to Reason?*

Casper Hansen, Christian Hansen, Lucas Chaves Lima. Automatic Fake News Detection: Are Models Learning to Reason? (To Appear) ACL-IJCNLP 2021: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [64].

Automatic Fake News Detection: Are Models Learning to Reason?

Casper Hansen
University of Copenhagen
c.hansen@di.ku.dk

Christian Hansen
University of Copenhagen
chrh@di.ku.dk

Lucas Chaves Lima
University of Copenhagen
lcl@di.ku.dk

Abstract

Most fact checking models for automatic fake news detection are based on reasoning: given a claim with associated evidence, the models aim to estimate the claim veracity based on the supporting or refuting content within the evidence. When these models perform well, it is generally assumed to be due to the models having learned to reason over the evidence with regards to the claim. In this paper, we investigate this assumption of reasoning, by exploring the relationship and importance of both claim and evidence. Surprisingly, we find on political fact checking datasets that most often the highest effectiveness is obtained by utilizing only the evidence, as the impact of including the claim is either negligible or harmful to the effectiveness. This highlights an important problem in what constitutes evidence in existing approaches for automatic fake news detection.

1 Introduction

Misinformation is spreading at increasing rates (Vosoughi et al., 2018), particularly online, and is considered a highly pressing issue by the World Economic Forum (Howell et al, 2013). To combat this problem, automatic fact checking, especially for estimating the veracity of potential fake news, have been extensively researched (Hassan et al., 2017; Hansen et al., 2019; Thorne and Vlachos, 2018; Elsayed et al., 2019; Allein et al., 2020; Popat et al., 2018; Augenstein et al., 2019). Given a claim, most fact checking systems are *evidence-based*, meaning they utilize external knowledge to determine the claim veracity. Such external knowledge may consist of previously fact checked claims (Shaar et al., 2020), but it typically consists of using the claim to query the web through a search API to retrieve relevant hits. While including the evidence in the model increases the effectiveness over using only the claim, existing work has not focused on the predictive power of isolated evidence, and hence whether it assists the model in enabling better reasoning.

In this work we investigate if fact checking models learn reasoning, i.e., provided a claim and associated evidence, whether the model determines the claim veracity by reasoning over the evidence. If the model learns reasoning, we would expect the following proposition to hold: *A model using both the claim and evidence should perform better on the task of fact checking compared to a model using only the claim or evidence*. If a model is only given the claim as input, it does not necessarily have the information needed to determine the veracity. Similarly, if the model is only given the evidence, the predictive signal must come from dataset bias or the differences in the evidence obtained from claims with varying veracity, as it otherwise corresponds to being able to provide an answer to an unknown question. In our experimental evaluation on two political fact checking datasets, across multiple types of claim and evidence representations, we find the evidence provides a very strong predictive signal independent of the claim, and that the best performance is most often obtained while entirely ignoring the claim. This highlights that fact checking models may not be learning to reason, but instead exploit an inherent signal in the evidence itself, which can be used to determine factuality independent of using the claim as part of the model input. This highlights an important problem in what constitutes evidence in existing approaches for automatic fake news detection. We make our code publicly available at https://github.com/casperhansen/fake-news-reasoning.

2 Related Work

Automatic fact checking models include deep learning approaches, based on contextual and non-contextual embeddings, which encode the claim

and evidence using RNNs or Transformers (Shaar et al., 2020; Elsayed et al., 2019; Allein et al., 2020; Popat et al., 2018; Augenstein et al., 2019; Hassan et al., 2017), and non-deep learning approaches (Wang, 2017; Pérez-Rosas et al., 2018), which uses hand-crafted features or bag-of-word representations as input to traditional machine learning classifiers such as random forests, SVM, and MLP (Mihalcea and Strapparava, 2009; Pérez-Rosas et al., 2018; Baly et al., 2018; Reddy et al., 2018).

Generally, models may learn to memorize artifact and biases rather than truly learning (Gururangan et al., 2018; Moosavi and Strube, 2017; Agrawal et al., 2016), e.g., from political individuals often leaning towards one side of the truth spectrum. Additionally, language models have been shown to implicitly store world knowledge (Roberts et al., 2020), which in principle could enhance the aforementioned biases. To this end, we design our experimental setup to include representative fact checking models of varying complexity (from simple term-frequency based representations to contextual embeddings), while always evaluating each trained model on multiple different datasets to determine generalizability.

3 Methods

Problem definition. In automatic fact checking of fake news we are provided with a dataset of $D = \{(c_1, e_1, y_1), ..., (c_n, e_n, y_n)\}$, where c_i corresponds to a textual claim, e_i is evidence used to support or refute the claim, and y_i is the associated truth label to be predicted based on the claim and evidence. Following current work on fact checking of fake news (Hassan et al., 2017; Thorne and Vlachos, 2018; Elsayed et al., 2019; Allein et al., 2020; Popat et al., 2018; Augenstein et al., 2019), we consider the evidence to be a list of top-10 search snippets as returned by Google search API when using the claim as the query. Note that while additional metadata may be available–such as speaker, checker, and tags–this work focuses specifically on whether models learn to reason based on the combination of claim and evidence, hence we keep the input representation to consist only of the latter.

Overview. In the following we describe the different models used for the experimental comparison (Section 4), which consists of models based on term frequency (term-frequency weighted bag-of-words (Salton and Buckley, 1988)), word embeddings (GloVe word embeddings (Pennington et al., 2014)), and contextual word embeddings (BERT (Devlin et al., 2019)). These representations are chosen as to include the typical representations most broadly used among past and current NLP models.

Term-frequency based Random Forest. We construct a term-frequency weighted bag-of-words representation per sample based on concatenating the text content of the claim and associated evidence snippets. We train a Random Forest (Breiman, 2001) as the classifier using the Gini impurity measure. In the setting of only using either the claim or evidence snippets as the input, only the relevant part is used for constructing the bag-of-words representation.

GloVe-based LSTM model. We adapt the model by Augenstein et al. (2019), which originally was proposed for multi-domain veracity prediction. Using a pretrained GloVe embedding (Pennington et al., 2014)[1], claim and snippet tokens are embedded into a joint space. We encode the claim and snippets using an attention-weighted bidirectional LSTM (Hochreiter and Schmidhuber, 1997):

$$h_{c_i} = \text{attn}\left(\text{BiLSTM}(c_i)\right) \quad (1)$$

$$h_{e_{i,j}} = \text{attn}\left(\text{BiLSTM}(e_{i,j})\right) \quad (2)$$

where attn$(\cdot)$ is a function that learns an attention score per element, which is normalized using a softmax, and returns a weighted sum. We combine the claim and snippet encodings using using the matching model by Mou et al. (2016) as:

$$s_{i,j} = \left[h_{c_i} \; ; \; h_{e_{i,j}} \; ; \; h_{c_i} - h_{e_{i,j}} \; ; \; h_{c_i} \cdot h_{e_{i,j}}\right] \quad (3)$$

where ";" denotes concatenation. The joint claim-evidence encodings are attention weighted and summed, projected through a fully connected layer into $\mathbb{R}^L$, where L is the number of possible labels:

$$o_i = \text{attn}([s_{i,1} \; ; \; ... \; ; \; s_{i,10}]) \quad (4)$$

$$p_i = \text{softmax}\left(\text{FC}(o_i)\right) \quad (5)$$

Lastly, the model is trained using cross entropy as the loss function. In the setting of using only the claim as the input (i.e., without the evidence), then h_{c_i} is used in Eq. 5 instead of o_i. If only the evidence is used, then an attention weighted sum of the evidence snippet encodings is used in Eq. 5 instead of o_i.

[1] http://nlp.stanford.edu/data/glove.840B.300d.zip

	Train: Snopes				Train: PolitiFact			
	Within dataset Eval: Snopes		Out-of dataset Eval: PolitiFact		Within dataset Eval: PolitiFact		Out-of dataset Eval: Snopes	
RF (∼13 seconds)	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$
Claim	0.473	0.231	**0.273**	0.223	0.254	0.255	0.546	0.243
Evidence	0.504	0.280	0.244	0.195	0.301	0.299	**0.597**	0.232
Claim+Evidence	0.550	0.271	0.245	0.190	0.310	0.304	0.579	0.207
LSTM (∼12 minutes, 888K parameters)								
Claim	0.408	0.243	0.260	**0.228**	0.237	0.237	0.565	0.221
Evidence	0.495	0.253	0.262	0.208	0.290	0.295	0.550	0.273
Claim+Evidence	0.529	0.253	0.258	0.189	0.288	0.294	0.509	0.256
BERT (∼264 minutes, 109M parameters)								
Claim	0.533	0.312	0.249	0.209	0.275	0.282	0.550	0.273
Evidence	0.531	**0.321**	0.249	0.224	**0.351**	**0.359**	0.577	**0.286**
Claim+Evidence	**0.556**	0.313	0.231	0.191	0.285	0.292	0.564	0.259

Table 1: Evaluation using micro and macro F1. Per column, the best score per method is underlined and the best score across all methods is highlighted in bold. We report the training time and number of model parameters, for Claim+Evidence on PolitiFact, in the parentheses. RF is trained on 5 cores and neural models on a Titan RTX.

BERT-based model. In a similar fashion to the LSTM model, we construct a model based on BERT (Devlin et al., 2019)[2], where the `[CLS]` token encoding is used for claim and evidence representations. Specifically, the claim and evidence snippets are encoded as:

$$h_{c_i} = \text{BERT}(c_i),\ h_{e_{i,j}} = \text{BERT}(c_i, e_{i,j}) \quad (6)$$

$$h_{e_i} = \text{attn}([h_{e_{i,1}}\ ;\ ...\ ;\ h_{e_{i,10}}]) \quad (7)$$

where the claim acts as the question when encoding the evidence snippets. Similarly to Eq. 5, the prediction is obtained by concatenating the claim and evidence representations and project it through a fully connected layer into $\mathbb{R}^L$:

$$p_i = \text{softmax}(FC([h_{c_i}\ ;\ h_{e_i}])) \quad (8)$$

where cross entropy is used as the loss function for training the model. In the setting that only the claim is used as input, then only h_{c_i} is used in Eq. 8. If only the evidence is used, then $h_{e_{i,j}}$ is computed without including c_i, and only h_{e_i} is used in Eq. 8.

4 Experimental Evaluation

4.1 Datasets

We focus on the domain of political fact checking, where we use claims and associated evidence from PolitiFact and Snopes, which we extract from the MultiFC dataset (Augenstein et al., 2019). Using the claim as a query, the evidence is crawled from Google search API as the search snippets of the top-10 results, and is filtered such that the website origin of a given claim does not appear as evidence. To facilitate better comparison between the datasets, we filter claims with non-veracity related labels[3]. The dataset statistics are shown in Table 2.

	#Claims	Labels
PolitiFact	13,581	pants on fire! (10.6%), false (19.2%), mostly false (17.0%), half-true (19.8%), mostly true (18.8%), true (14.8%)
Snopes	5,069	false (64.3%), mostly false (7.5%), mixture (12.3%), mostly true (2.8%), true (13.0%)

Table 2: Dataset statistics.

4.2 Experimental setup

Both datasets are split into train/val/test sets using label-stratified sampling (70/10/20% splits). We tune all models on the validation split, and use early stopping with a patience of 10 for neural models. Following Augenstein et al. (2019), we use micro and macro F1 for evaluation. The models are evaluated on both the within dataset test sets, but also out-of dataset test sets (e.g., a model trained on Snopes is evaluated on both Snopes and PolitiFact).

[2]We use `bert-base-uncased` from https://huggingface.co/bert-base-uncased.

[3]For PolitiFact we exclude [full flop, half flip, no flip] and for Snopes we exclude [unproven, miscaptioned, legend, outdated, misattributed, scam, correct attribution].

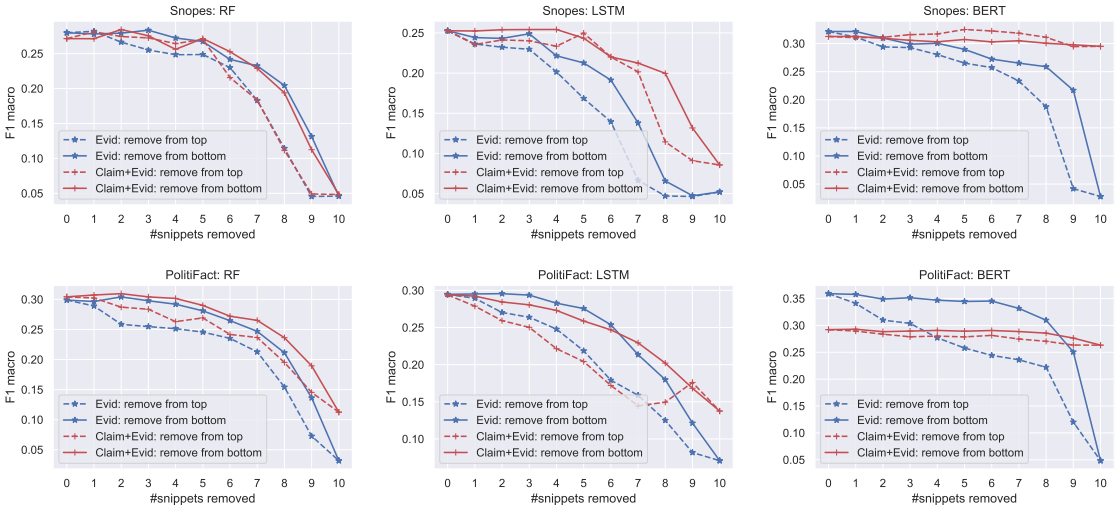


Figure 1: Macro F1 scores when removing evidence from either the top or bottom of the evidence snippet ranking.

In the out-of dataset evaluation we need the labels to be comparable, hence in that setting we merge "pants on fire!" and "false" for PolitiFact.

5 Tuning details

In the following, the best overall parameter configurations are underlined. The best configuration is chosen based on the average of the micro and macro F1[4]. For RF, we tune the number of trees from [100,500,1000], the minimum number of samples in a leaf from [1,3,5,10], and the minimum number of samples per split from [2,5,10]. For the LSTM model, we tune the learning rate from [1e-4,5e-4,1e-5], batch size [16,32], number of LSTM layers from [1,2], dropout from [0, 0.1], and fix the number of hidden dimensions to 128. For the BERT model, we tune the learning rate from [3e-5, 3e-6, 3e-7] and fix the batch size to 8.

5.1 Results

The results can be seen in Table 1. Overall, we see that the BERT model trained only on Evidence obtains the best results in 4/8 columns, and, notably, in 3/4 cases the BERT model with Evidence obtains the best macro F1 score on within and out-of dataset prediction. Random forest using term-frequency as input obtains the best out-of dataset micro F1 for both datasets (using either only Claim or only Evidence). Across all methods, the combination of Claim+Evidence only marginally obtains the best results a single time (for Snopes micro F1). For further details, in Table 3 we compute the accuracy scores for all the false labels, mixture or half-true label, and true labels.

Surprisingly, a BERT model using only the Evidence is capable of predicting the veracity of the claim used for obtaining the evidence. This shows that a strong signal must exist in the evidence itself, and the evidence found by the search engine appears to be implicitly affected by the veracity of the claim used as the query in some way[5]. The improvements reported in the literature by combining claim and evidence, are therefore not evident of the model learning to reason over the evidence with regards to the claim, but instead exploiting a signal inherent in the evidence itself. This highlights that the current approach for evidence gathering is problematic, as the strong signal makes it possible (and most often beneficial) for the model to entirely ignore the claim. This makes the model entirely reliant on the process behind how the evidence is generated, which is outside the scope of the model, and thereby undesirable, as any change in the search system may change the model performance significantly. It may also be problematic on a more fundamental level, e.g., to predict the veracity of the following two claims: "the earth is round" and "the earth is flat", the evidence could be the same, but a model entirely dependent on the evidence, and not the claim, would be incapable of predicting both claims correctly.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[5] Note that the claim origin website is always removed from the evidence.

	Train: Snopes						Train: PolitiFact					
	Within dataset Eval: Snopes			Out-of dataset Eval: PolitiFact			Within dataset Eval: PolitiFact			Out-of dataset Eval: Snopes		
RF	acc_{false}	acc_{mix}	acc_{true}	acc_{false}	acc_{mix}	acc_{true}	acc_{false}	acc_{mix}	acc_{true}	acc_{false}	acc_{mix}	acc_{true}
Claim	0.710	0.144	0.255	0.853	0.016	0.209	0.623	0.216	0.513	0.790	0.092	0.255
Evidence	0.705	0.152	0.441	0.829	0.006	0.117	0.654	0.248	0.510	**0.891**	0.039	0.192
Claim+Evidence	**0.760**	0.136	0.453	0.829	0.000	0.117	0.634	0.292	0.512	0.871	0.039	0.199
LSTM												
Claim	0.674	0.232	0.280	0.875	0.047	0.137	0.566	0.212	0.504	0.833	0.026	0.234
Evidence	0.721	0.272	0.267	**0.890**	0.020	0.115	0.643	0.253	0.485	0.768	0.184	0.322
Claim+Evidence	0.757	0.248	0.168	0.879	0.008	0.107	**0.671**	0.210	0.460	0.704	0.171	**0.378**
BERT												
Claim	0.746	0.256	0.379	0.854	**0.094**	0.045	0.604	0.292	0.475	0.765	0.171	0.287
Evidence	0.648	**0.376**	**0.559**	0.702	0.049	**0.337**	0.649	**0.326**	0.496	0.804	**0.197**	0.339
Claim+Evidence	0.747	0.264	0.379	0.882	0.067	0.042	0.667	0.175	**0.558**	0.790	0.092	0.367

Table 3: Accuracy scores computed on the false labels, mixture or half-true label, and true labels. All labels within a group (e.g., any false label such as false or mostly false) are considered to be the same and as such this reduces the problem to a three class classification problem.

5.2 Removal of evidence

We observed a strong predictive signal in the evidence alone and now consider the performance impact when gradually removing evidence snippets. The evidence is removed consecutively either from the top down or bottom up (i.e., removing the most relevant snippets first and vice versa), until no evidence is used. Figure 1 shows the macro F1 as a function of removed evidence when using the Evidence or Claim+Evidence models. We observe a distinct difference between the random forest and LSTM model compared to BERT: for random forest and LSTM, the Claim+Evidence models on both datasets drop rapidly in performance when the evidence is removed, while the BERT model only experiences a very small drop. This shows that when the Claim+Evidence is used in the BERT model, the influence of the evidence is minimal, while the evidence is vital for the Claim+Evidence RF and LSTM models. For all models, we observe that when evidence is removed from the top down, the performance drop is larger than when evidence is removed from the bottom up. Thus, the ranking of the evidence as provided by the search engine is related to its usefulness as evidence for fact checking.

6 Conclusion

We investigate if fact checking models for fake news detection are learning to process claim and evidence jointly in a way resembling reasoning. Across models of varying complexity and evaluated on multiple datasets, we find that the best performance can most often be obtained using only the evidence. This highlights that models using both claim and evidence are inherently not learning to reason, and points to a potential problem in how evidence is currently obtained in existing approaches for automatic fake news detection.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Lee Howell et al. 2013. Digital wildfires in a hyperconnected world. *WEF report*, 45(3):15–94.

Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. 2020. Time-aware evidence ranking for fact-checking. *arXiv preprint arXiv:2009.06402*.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4677–4691.

Ramy Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pages 21–27.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019. Checkthat! at clef 2019: Automatic identification and verification of claims. In *Advances in Information Retrieval*, pages 309–315, Cham. Springer International Publishing.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. 2019. Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 994–1000.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, page 309–312, USA. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 130–136.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. 2018. Defactonlp: Fact verification using entity recognition, TFIDF vector comparison and decomposable attention. *CoRR*, abs/1809.00509.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426. Association for Computational Linguistics.

9. *Principled Multi-Aspect Evaluation Measures of Rankings.*

Maria Maistro, Lucas Chaves Lima, Jakob Grue Simonsen, Christina Lioma. Principled Multi-Aspect Evaluation Measures of Rankings (Under revision). 2021, 30th ACM International Conference on Information and Knowledge Management (CIKM2021) [111]

Principled Multi-Aspect Evaluation Measures of Rankings

Anonymous Author(s)

ABSTRACT

Information Retrieval evaluation has traditionally focused on defining principled ways of assessing the relevance of a ranked list of documents with respect to a query. Several methods extend this type of evaluation beyond relevance, making it possible to evaluate different aspects of a document ranking (e.g., relevance, usefulness, or credibility) using a single measure (*multi-aspect evaluation*). However, these methods either are (i) tailor-made for specific aspects and do not extend to other types or numbers of aspects, or (ii) have theoretical anomalies, e.g. assign maximum score to a ranking where all documents are labelled with the lowest grade with respect to all aspects (e.g., not relevant, not credible, etc.).

We present a theoretically principled multi-aspect evaluation method that can be used for any number, and any type, of aspects. A thorough empirical evaluation using up to 5 aspects and a total of 425 runs officially submitted to 10 TREC tracks shows that our method is more discriminative than the state-of-the-art and overcomes theoretical limitations of the state-of-the-art.

KEYWORDS

Evaluation, ranking, multiple aspects, partial orders

ACM Reference Format:

Anonymous Author(s). 2021. Principled Multi-Aspect Evaluation Measures of Rankings. In *CIKM '21: 30th ACM International Conference on Information and Knowledge Management, November 1–5, 2021, Gold Coast, Queensland, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Multi-aspect evaluation is a task in *Information Retrieval (IR)* evaluation where the ranked list of documents returned by an IR system in response to a query is assessed in terms of not only relevance, but also other *aspects* (or dimensions) of the ranked documents, such as credibility or usefulness. Generally, there are two ways to conduct multi-aspect evaluation: (1) evaluate each aspect separately using any appropriate single-aspect evaluation measure (e.g., AP, NDCG, F1), and then aggregate the scores across all aspects into a single score; or (2) evaluate all aspects at the same time using any appropriate multi-aspect evaluation measure [5, 35, 42]. An advantage of the aggregating option (1) is that it is easy to implement using evaluation measures that are readily available and well-understood in the community. Its disadvantage is that it is not guaranteed that all aspects will have similar distributions of labels, and aggregating across wildly different distributions can give odd results [34].

The second way of doing multi-aspect evaluation is to use a single multi-aspect evaluation measure. The problem here is that few such evaluation measures exist, and most of them are defined for specific aspects and do not generalise to other types/numbers of aspects (see an exhaustive review of these measures in §2).

Motivated by the above, we **contribute** a novel multi-aspect evaluation method that works with any type and number of aspects, and avoids the above problems. Given a ranked list of documents, where documents are labelled with respect to multiple aspects, our method, *Total Order Multi-Aspect (TOMA)* evaluation, first defines a preferential order (formally *weak total order relation*) between the documents with respect to their multiple aspect labels, and then aggregates the document labels across multiple aspects to obtain a ranking of aggregated aspect labels, which can be evaluated by any single-aspect evaluation measure, such as *Normalized Discounted Cumulated Gain (NDCG)* or *Average Precision (AP)*. Simply put, instead of evaluating each aspect separately and then aggregating their scores, we first aggregate the aspect labels and then evaluate the ranked list of documents. We do this in a way that provides several degrees of freedom: our method can be used with any number and type of aspects, can be instantiated with any binary or graded, set-based or rank-based evaluation measure, and can accommodate any granularity in the importance of each aspect or label, but still ensures, by definition, that the preference order among multi-aspect documents is not violated, and that the final measure score will meet some common requirements, i.e., the minimum (worst) score being 0 and the maximum (perfect) score being 1. We validate this empirically (§4) and theoretically (§4.2).

2 RELATED WORK

Multi-aspect evaluation measures for IR have been studied for different tasks and aspects, starting from the INEX initiative with a discussion on relevance and coverage [30]. Since then, measures have been proposed to evaluate relevance and novelty or diversity, such as *α-Normalized Discounted Cumulated Gain (α-NDCG)* [16], *Intent Aware Mean Average Precision (MAP-IA)* [1] and *Intent Aware Expected Reciprocal Rank (IA-ERR)* [10]; relevance, novelty and the amount of user effort, such as *Normalized Cube Test (nCT)* [42]; relevance, redundancy and user effort, such as *Rank-Biased Utility (RBU)* [5]; relevance and understandability, such as *Understandability-biased Rank-Biased Precision (uRBP)* [49] and the *Multidimensional Measure (MM)* framework [37]; and relevance and credibility, such as *Normalised Local Rank Error (NLRE)*, *Normalised Global Rank Error (NGRE)*, *Normalised Weighted Cumulative Score (nWCS)*, *Convex Aggregating Measure (CAM)* and the *Weighted Harmonic Mean Aggregating Measure (WHAM)* [35]. All these measures have limitations; we describe these next.

Firstly, except for RBU, none of the above measures are based on a formal framework. They are defined as stand-alone tools to assess the effectiveness of a ranked list of documents. This means

CIKM '21, November 1–5, 2021, Gold Coast, Queensland, Australia

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

that, even if the measure can assess the effectiveness of an input ranking, the order induced by the measure over the space of input rankings is not well-defined. Hence, there is no canonical *ideal ranking* [1] that is well-defined or easy to compute, e.g., for α-NDCG, the computation of the ideal ranking is equivalent to a minimal vertex covering problem [16], an NP-complete problem, while for CT and nCT, computing the ideal ranking is equivalent to the minimum edge dominating set problem [42], an NP-hard problem. Computationally better ways of comparing to an ideal ranking can be devised using graded similarity—so-called *effectiveness levels* to an ideal ranking using rank-biased overlap [17–19], but this approach requires defining a (set of) ideal ranking(s), which has not appeared for multi-aspect ranking prior to the present paper.

Evaluation measures that do not compare against an ideal ranking may be harder to interpret or problematic. *Discounted Cumulated Gain (DCG)* is not upper bounded by 1, thus different topics are not weighted equally and scores are not comparable. Failing to compare against the ideal ranking is problematic in multi-aspect evaluation: α-NDCG allows systems to reach scores greater than 1, which is supposed to be the score of the perfect system. With NLRE and NGRE, a system that retrieves no relevant or credible documents has error =0, i.e., achieves the best score, because the relative order of pairs of documents is always correct [34]. Similarly, nWCS can reach the perfect score of 1, even if no relevant or credible documents are retrieved, since the normalization is computed with a re-ranking of the input ranking, instead of the ideal ranking.

Both uRBP and RBU have a different problem: to reach the perfect score of 1, a system must retrieve an infinite number of relevant and understandable documents, even if those documents are not available in the collection. CAM and WHAM use the weighted arithmetic and weighted harmonic mean of any IR measure computed with respect to relevance and credibility independently. Therefore, depending on the distribution of labels across the aspects, it can be impossible for any system to reach the perfect score (see § 4.2).

Secondly, most of the above multi-aspect evaluation measures are defined for specific contexts and with a limited set of aspects, e.g., novelty, diversity, credibility and understandability, thus they cannot deal with a more general scenario and a variable number of aspects. For RBU, even though a formal framework is defined, its formulation specifies only diversity and redundancy constraints, which cannot be applied to a general set of aspects. This inability to generalise to more/other types of aspects means that, if a system must be evaluated with respect to a new aspect, the measure needs to be properly adapted. This can be easily done for some measures, e.g., CAM, WHAM, and nWCS, but the lack of a formal framework behind them may lead to odd results, e.g., extending NLRE to 3 aspects returns a score distribution compressed towards 0, preventing the rankings to be evaluated in a fair way [34].

3 TOMA FRAMEWORK

We formalize the problem and our proposed methodology: we explain why reasoning in terms of multiple aspects leads to a partial order relation among documents (§ 3.1); how we complete the partial order relation with the distance order (§ 3.2); and how to use the distance order with state-of-the-art IR evaluation measures (§ 3.3).

[1] An *ideal ranking* is the best ranking of all assessed documents for a given topic [29].

3.1 Formalization of the Problem

Let $A = \{a_1, \ldots, a_n\}$ be a set of *aspects*; each aspect $a \in A$ has a non-empty set of *labels* $L_a = \{l^a_0, \ldots, l^a_{K_a}\}$ and an order relation $\prec_a$ such that: $l^a_0 \prec_a l^a_1 \prec_a \cdots \prec_a l^a_{K_a}$, e.g., we may have 2 aspects $A = \{\text{relevance}, \text{correctness}\}$, with the set $L_r = \{\mathtt{nr}, \mathtt{mr}, \mathtt{fr}, \mathtt{hr}\}$ (non-relevant, marginally relevant, fairly relevant, highly relevant) ordered as: $\mathtt{nr} \prec_r \mathtt{mr} \prec_r \mathtt{fr} \prec_r \mathtt{hr}$; and the set $L_c = \{\mathtt{nc}, \mathtt{pc}, \mathtt{c}\}$ (non-correct, partially correct, correct) ordered as: $\mathtt{nc} \prec_c \mathtt{pc} \prec_c \mathtt{c}$. Let D be the set of *documents* and T the set of *topics*. Each document $d \in D$ is mapped to a *ground truth* vector $\mathrm{GT}(d, t) = (l_1, \ldots, l_n) \in L_{a_1} \times \cdots \times L_{a_n}$ that contains the "true" label of d for each aspect, e.g., a document may have $\mathrm{GT}(d, t) = (\text{highly relevant}, \text{non-correct})$.

In IR, given a topic t, the objective is to rank documents in D such that for the documents $d', d \in D$, if d' is ranked before d, then $\mathrm{GT}(d, t) \preceq_* \mathrm{GT}(d', t)$ for a given order relation $\preceq_*$. When there is only one aspect $A = \{a\}$, one can use $\prec_a$, the order on the set of labels L_a, to induce a weak order on D and decide if d' should be ranked before d. If only relevance is assessed, we consider the relation induced by relevance labels, i.e., documents labelled "highly relevant" should be ranked before "fairly/marginally relevant" and "non-relevant" documents. Applying this approach to multiple aspects requires reasoning about orderings of tuples of labels with different aspects, e.g., for documents $d', d \in D$, such that $\mathrm{GT}(d', t) = (\text{highly relevant}, \text{correct})$ and $\mathrm{GT}(d, t) = (\text{marginally relevant}, \text{correct})$, it is reasonable to rank d' before d.

Indeed, there is one *unequivocal* way of deeming one document better than another, and this is if document d' has better labels than document d for *every* aspect: if for $\mathrm{GT}(d, t) = (l_1, \ldots, l_n)$ and $\mathrm{GT}(d', t) = (l'_1, \ldots, l'_n)$ we have $l_i \preceq_{a_i} l'_i$ for all $i \in \{1, \ldots, n\}$, then any document labeled $(l'_1, \ldots, l'_n)$ is better or equal than any document labelled $(l_1, \ldots, l_n)$ and should occur before it in a "good" ranking. We denote this order relation by $\mathrm{GT}(d, t) \sqsubseteq \mathrm{GT}(d', t)$.

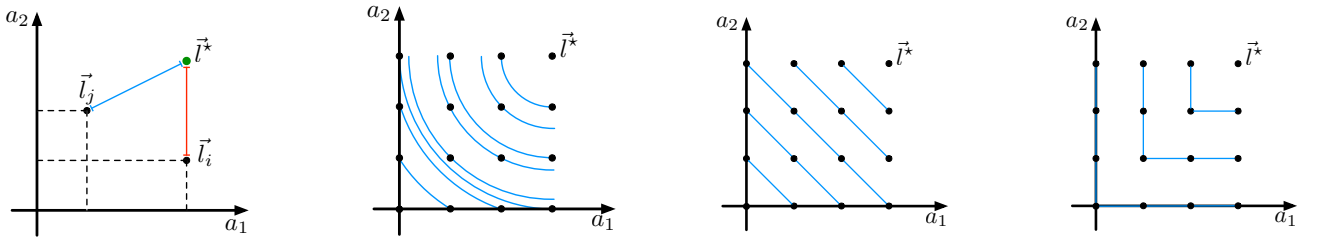
The order relation $\sqsubseteq$ leads to a *partial* instead of a *total* order, i.e., there are documents that are *not comparable*[2], e.g., if d' is now highly relevant and partially correct, the final ranking is not clear: should one promote d' (more relevant) or d (more correct)? This is an example of documents that are not comparable, so we have $\mathrm{GT}(d, t) \not\sqsubseteq \mathrm{GT}(d', t)$ and $\mathrm{GT}(d', t) \not\sqsubseteq \mathrm{GT}(d, t)$, and the choice of whether d' is preferred to d may lie on the intended application.

A partial order relation and the presence of not comparable documents implies that it is not possible to univocally rank the documents in D. If we could "complete" the partial order with a total order, or at least a weak order, we could rank documents and define an ideal ranking, where for any $d', d \in D$, the order relation determines the rank position of d' and d. So, before tackling the problem of evaluating a ranked list of documents in a multi-aspect way, we build such an order relation. This is detailed next.

3.2 The distance order

We now explain how to obtain a weak order relation from the partial order relation $\sqsubseteq$. Consider the Cartesian product of all sets of labels $L = L_{a_1} \times \cdots \times L_{a_n}$. An element $\mathbf{l} \in L$ is a tuple of labels $\mathbf{l} = (l_1, \ldots, l_n)$. The total order relation will be denoted by $\preceq_*$ and it

[2] A partial order is reflexive, antisymmetric and transitive; a total order is a partial order where all pairs of items are comparable; a weak order is a total order without antisymmetry [28].



(a) Distance order. (b) Euclidean distance. (c) Manhattan distance. (d) Chebyshev distance.

Figure 1: Example with two aspects a_1 and a_2. Each point is a tuple of labels. The best label $\mathbf{l}^\star$ is in the top right. The distance between tuples of labels and $\mathbf{l}^\star$ defines a total order relation. Blue lines connect tuples of labels at the same distance from $\mathbf{l}^\star$.

will be a weak order relation on L, i.e., a total binary relation that is reflexive and transitive, but not necessarily anti-symmetric [25–27]. This weak order allows *all* tuples of labels to be compared, i.e., for any two $\mathbf{l}, \mathbf{l}' \in L$ we will have $\mathbf{l}' \preceq_* \mathbf{l}$ and/or $\mathbf{l} \preceq_* \mathbf{l}'$. Consequently, all documents will be comparable through their tuple of labels.

We require that the weak order relation $\preceq_*$ respects the partial order relation $\sqsubseteq$:

$$\forall\, \mathbf{l}, \mathbf{l}' \in L \text{ we have } \mathbf{l} \sqsubseteq \mathbf{l}' \Rightarrow \mathbf{l} \preceq_* \mathbf{l}' \tag{1}$$

This means that, for comparable documents, the partial order relation and the weak order relation rank documents in the same way. Moreover the weak order relation allows to rank even those documents that are not comparable with the partial order relation.

To define $\preceq_*$, we embed the tuples of labels in the Euclidean space and derive the weak order $\preceq_*$ using known distance functions. Let g be an embedding function that maps tuples of labels in Euclidean space $\mathcal{L} = \mathbb{R}^n$: $g(\mathbf{l}) = g(l_1, \ldots, l_n) = (g_{a_1}(l_1), \ldots, g_{a_n}(l_n))$. We assume that for each $a \in A$, g_a is a non-decreasing map, i.e., for any $l, l' \in L_a$ if $l \preceq_a l'$ then $g_a(l) \leq g_a(l')$. Intuitively, g_a assigns a number to each label, which allows to represent tuples of labels in the Euclidean space. We illustrate in §3.4 how the embedding function g affects the final ranking of documents.

Through the embedding function g, each tuple of labels $\mathbf{l}$ is represented by a point in the Euclidean space $\mathcal{L}$ denoted by $\vec{l} = g(\mathbf{l})$. We define the *best label tuple* as the tuple of labels $\mathbf{l}^\star$ whose coordinates are the best label for each aspect, $\mathbf{l}^\star = (l_{K_{a_1}}, \ldots, l_{K_{a_n}})$. The idea is to treat $\mathbf{l}^\star$ as the maximum element and use the distance from this maximum element to define the desired weak order relation; e.g., for two aspects a_1 and a_2, each tuple of labels is represented as a point in the Euclidean plane, and the best label $\mathbf{l}^\star$ is represented by the topmost and right-most point (see Fig. 1a). Then, given two documents d and d', d is ranked before d' if $\mathrm{GT}(d, t)$ is closer to the best label than $\mathrm{GT}(d', t)$.

We formally define the *distance order* as the following relation:

$$\mathbf{l} \preceq_* \mathbf{l}' \iff \mathrm{Dist}(\vec{l}, \vec{l}^\star) \geq \mathrm{Dist}(\vec{l}', \vec{l}^\star) \tag{2}$$

where $\mathrm{Dist}\colon \mathcal{L} \times \mathcal{L} \to [0, +\infty[$ is any function such that $\mathrm{Dist}(\vec{l}^\star, \vec{l}^\star) = 0$[3]. The relation $\preceq_*$ is a weak order: all $\mathbf{l}, \mathbf{l}'$ are comparable because $\mathrm{Dist}(\vec{l}, \vec{l}^\star)$ is defined for all $\mathbf{l}$, and as $\geq$ is reflexive and transitive on $[0, +\infty[$, the relation $\preceq_*$ is reflexive and transitive (but not necessarily antisymmetric). Since the distance order is a weak order, it allows to deem items "equally good" when it is impossible or undesirable to impose a strict total order[4]. Thus we write:

$$\mathbf{l} =_* \mathbf{l}' \iff \mathrm{Dist}(\vec{l}, \vec{l}^\star) = \mathrm{Dist}(\vec{l}', \vec{l}^\star) \tag{3}$$

which means that $\vec{l}$ and $\vec{l}'$ are at the same distance from $\vec{l}^\star$.

Note that the distance order can be *tailored*: we may instantiate Dist with any valid distance function. We illustrate this in Fig. 1b-1d with *Euclidean* (order relation $\preceq_2$), *Manhattan* (order relation $\preceq_1$), and *Chebyshev* (order relation $\preceq_\infty$). With these choices of Dist, the distance order defined in Eq. (2)-(3) respects the partial order $\sqsubseteq$, which means that it satisfies the requirement in Eq. (1) because g_a is a non decreasing map (a proof is provided in the online appendix[5]).

3.3 Integration with IR measures

Next we integrate the distance order with known IR measures such as AP or NDCG. The binary relation $=_*$ in Eq. (3) is an equivalence relation. Given a tuple of labels $\mathbf{l} \in L$, its equivalence class $[\mathbf{l}]_*$ is the set of all tuples of labels with equal distance from the best label $[\mathbf{l}]_* = \{\mathbf{l}' \in L\colon \mathrm{Dist}(\vec{l}', \vec{l}^\star) = \mathrm{Dist}(\vec{l}, \vec{l}^\star)\}$.

Inducing the relation defined in Eq. (2) on the set of documents D allows to rank documents by their membership to each equivalence class, which corresponds to the distance of their tuple of labels to the best label. We place closest to the top of the ranking documents whose equivalence class is closest to the best label, and vice versa.

To combine the distance order with IR measures we map each equivalence class (set of tuple of labels), to a non negative integer. This is similar to what happens with single-aspect evaluation, where each label is mapped to a weight: e.g., with 4 relevance labels, we can compute NDCG with equi-spaced relevance weights $\{0, 1, 2, 3\}$ [29] or exponential weights $\{0, 2, 4, 8\}$ [8]. We define a *weight function* $W\colon L \to \mathbb{N}_0^+$ as a map such that the order relation $\preceq_*$ is preserved:

$$\forall\, \mathbf{l}, \mathbf{l}' \in L\colon \mathbf{l} \preceq_* \mathbf{l}' \implies W(\mathbf{l}) \leq W(\mathbf{l}') \tag{4}$$

where the constraint in Eq. (4) entails that W is a non-decreasing function with respect to the weak order $\preceq_*$ on the set of tuples of labels. This means that W can return different integers for each equivalence class, but also the same integer for different equivalence

[3] Distance functions must be symmetric and satisfy the triangle inequality. Any such distance function satisfies our condition on Dist, and so do our example distances.

[4] This is the reason that *weak* orders (that are not necessarily anti-symmetric), rather than strict total orders, are typically used in the literature [27, 47].

[5] https://mega.nz/file/mygEDBQC#zslzdU7lBLrACyMitAb2YmOwNowQP-YXvFOuHD82jn4

classes, i.e., 0 and 1, whenever we need to compute a binary single-aspect IR measure as AP.

To summarize, our TOMA method has 3 steps:

(1) We embed tuples of labels into elements of Euclidean space, and we derive the weak order $\preceq_*$ using a distance function;
(2) We define an adjustable weight function W that preserves $\preceq_*$ and maps each tuple of labels to a single integer weight (this allows to aggregate tuple of labels so that better documents can be given greater weight);
(3) Having such a weak order and the weight function W, any existing single-aspect IR evaluation measure can be used to assess the quality. Thus, we choose a single-aspect evaluation measure μ and compute the final evaluation score as $M = \mu \circ W : M(r_t) = \mu(W(\mathrm{GT}(d_1, t)), \ldots, W(\mathrm{GT}(d_N, t)))$, where r_t is a ranked list of documents.

The above is compatible with any number and type of aspect.

3.4 Example

We present an example on the role of different choices of embedding functions, distance functions and weight functions in TOMA with 4 relevance labels {nr, mr, fr, hr} and 3 correctness labels {nc, pc, c}. As in real scenarios [34], we assume that not relevant documents are not correct: as they do not include information about the topic, they cannot be correct with respect to that topic.

Tab. 1 shows 3 different embeddings for correctness; the embedding for relevance is fixed. Note that the distance functions are invariant under translations and rotations, thus, rather than the actual values assigned from the embedding function g, it is important to consider the relation between different aspects. Independently of the choice of the embedding function and due to the definition of the selected distance functions, we see that: (i) Chebyshev generates the least number of equivalence classes and deems many tuples of labels as equal, since by taking the maximum it considers just the "furthest" or worst aspect to compute the distance; (ii) Manhatthan is somehow in-between Chebyshev and Euclidean and generates the equivalence classes by taking the sum across aspects; (iii) Euclidean generates the highest number of equivalence classes as it differentiates among tuples more than Manhatthan and is more sensitive to extreme cases, e.g., cases where one aspect has the best label and all other aspects have the lowest label.

In the 1^{st} scenario of Tab. 1 we map relevance and correctness to the same interval $[0, 3]$ (i.e., a highly relevant document is as "important" as a correct document). All labels are equi-spaced in the given range (the difference between a fairly relevant and a marginally relevant document is the same as that between a highly relevant and a fairly relevant one). In the case of fake news, with the Euclidean distance all relevant and not correct documents will be deemed worse than all other documents, but will be placed before not relevant and not correct documents. On the other hand, Chebyshev places relevant and not correct documents in the same equivalence class as not relevant documents, so those documents do not provide any contribution and can be simply filtered out. Manhattan represents a middle solution: highly relevant and not correct documents are deemed better than marginally relevant and partially correct documents, but worse than all other correct or partially correct documents.

In the 2^{nd} scenario of Tab. 1 relevance and correctness are mapped to different ranges, but all labels are equi-spaced with the same step of size 1. Here, relevance is more important than correctness. This is reflected on the sorting of equivalence classes: for all distance functions, highly relevant and not correct documents do not belong to the worst equivalence classes, but they are somehow better than partially correct documents. Even Chebyshev, which can be seen as the "strictest" distance function, places all relevant and not correct documents in the same equivalence class, which is considered better than the equivalence class of not relevant and not correct documents.

In the 3^{rd} scenario of Tab. 1 correctness is mapped to a range twice the size as the relevance range and we do not use equi-spaced labels for correctness. We assign more importance to correctness than relevance, and among correctness labels we penalize not correct and partially correct documents. The result is that for all distance functions relevant and not correct documents are considered among the worst equivalence classes. This particular setting affects also the other equivalence classes: correctness is preferred over relevance, e.g., correct documents should be always ranked before partially correct documents, regardless of their relevance label.

Note that TOMA requires a weight function satisfying the requirement in Eq. (4). If we wish to reward a system for sorting documents exactly as presented by the equivalence classes in Tab. 1, then the weight function should assign a different integer to each equivalence class. This choice of weight is similar to the choice of weights for relevance labels and its impact on the evaluation outcome is strictly tight to the evaluation measure used, as for example when one considers NDCG with different weighting schemes [29].

4 EXPERIMENTAL EVALUATION

We evaluate TOMA on 425 rankings that were submitted as official runs to 10 TREC tracks [11–15, 20, 21, 34, 45, 48] (see Tab. 2).

4.1 Experimental Setup

We use up to 5 different aspects. All aspects are assessed by TREC assessors as part of the corresponding track, except *popularity* and *non-spamminess*. We approximate *popularity* by PageRank[6], and *non-spamminess* by the Waterloo Spam Ranking[7]. We discretize the PageRank scores to generate 3 grades of popularity (`not popular`, `fairly popular`, `highly popular`), while simulating a power law distribution of popular and not popular documents (few highly popular (5%), some fairly popular (10%), and most not popular (85%)). For non-spamminess, we generate 3 grades of labels (`spam`, `fairly spam`, `not spam`) from the Waterloo Spam Ranking. We treat any document with score below 80 as spam (77%), documents with score in [80, 89] (14%) as `fairly spam`, and documents with score greater than 90 (9%) as `not spam` [38].

For the Web 2010-2014 and Task 2015-2016 tracks, we merge the labels `junk` and `non relevant` into `non relevant`, as was done by the TREC track organisers. For Task 2015-2016, Decision 2019 and Misinformation 2020, *usefulness*, *credibility*, and *correctness* were not assessed for not relevant documents, thus not relevant documents are assumed to be not useful, not credible, and not correct.

[6] http://www.lemurproject.org/clueweb12/PageRank.php
[7] https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/

Table 1: Final ordering of tuples of labels embedded in the Euclidean space. Relevance labels are always embedded in the same mapping (under *Relevance*). We use different mappings for correctness labels (under *Correctness*). Tuples that are relevant and not correct (high-traffic fake news) are in red.

Relevance	Correctness	Distance	Order among Tuples of Labels
$\{0,1,2,3\}$	$\{0,3/2,3\}$	Euclidean	$(3,3) \preceq_* (2,3) \preceq_* (3,3/2) \preceq_* (2,3/2) \preceq_* (1,3) \preceq_* (1,3/2) \preceq_* (3,0) \preceq_* (2,0) \preceq_* (1,0) \preceq_* (0,0)$
		Manhattan	$(3,3) \preceq_* (2,3) \preceq_* (3,3/2) \preceq_* (1,3) \preceq_* (2,3/2) \preceq_* (3,0) \preceq_* (1,3/2) \preceq_* (2,0) \preceq_* (1,0) \preceq_* (0,0)$
		Chebyshev	$(3,3) \preceq_* (2,3) \preceq_* (3,3/2) =_* (2,3/2) \preceq_* (1,3) =_* (1,3/2) \preceq_* (3,0) =_* (2,0) =_* (1,0) =_* (0,0)$
$\{0,1,2,3\}$	$\{0,1,2\}$	Euclidean	$(3,2) \preceq_* (3,1) =_* (2,2) \preceq_* (2,1) \preceq_* (3,0) =_* (1,2) \preceq_* (2,0) =_* (1,1) \preceq_* (1,0) \preceq_* (0,0)$
		Manhattan	$(3,2) \preceq_* (3,1) =_* (2,2) \preceq_* (3,0) =_* (2,1) =_* (1,2) \preceq_* (2,0) =_* (1,1) \preceq_* (1,0) \preceq_* (0,0)$
		Chebyshev	$(3,2) \preceq_* (3,1) =_* (2,1) =_* (2,2) \preceq_* (3,0) =_* (2,0) =_* (1,0) =_* (1,1) =_* (1,2) \preceq_* (0,0)$
$\{0,1,2,3\}$	$\{0,2,6\}$	Euclidean	$(3,6) \preceq_* (2,6) \preceq_* (1,6) \preceq_* (3,2) \preceq_* (2,2) \preceq_* (1,2) \preceq_* (3,0) \preceq_* (2,0) \preceq_* (1,0) \preceq_* (0,0)$
		Manhattan	$(3,6) \preceq_* (2,6) \preceq_* (1,6) \preceq_* (3,2) \preceq_* (2,2) \preceq_* (1,2) =_* (3,0) \preceq_* (2,0) \preceq_* (1,0) \preceq_* (0,0)$
		Chebyshev	$(3,6) \preceq_* (2,6) \preceq_* (1,6) \preceq_* (3,2) =_* (2,2) =_* (1,2) \preceq_* (3,0) =_* (2,0) =_* (1,0) =_* (0,0)$

Table 2: Experimental data. All aspects are labelled by TREC except popularity († approximated by PageRank) and non-spamminess (‡ approximated by Waterloo Spam Ranking). * means that the `junk` labels are merged with `non relevant`.

	TREC tracks									
	Web 2009	**Web 2010**	**Web 2011**	**Web 2012**	**Web 2013**	**Web 2014**	**Task 2015**	**Task 2016**	**Decision 2019**	**Misinfo2020**
Collection	ClueWeb09				ClueWeb12				ClueWeb12-B13	CommonCrawl News
Topics	50	48	50	50	50	50	35	50	50	46
Submitted runs	71	56	61	48	61	30	6	9	32	51
Aspects (label grades)	relevance (4) popularity† (3) non-spam‡ (3)	relevance (5*) popularity† (3) non-spam‡ (3)	relevance (4*) popularity† (3) non-spam‡ (3)	relevance (5*) popularity† (3) non-spam‡ (3)			relevance (3*) usefulness (3) popularity† (3) non-spam‡ (3)		relevance (3) credibility (2) correctness (2)	relevance (2) credibility (2) correctness (2)

We evaluate three versions of our method, TOMA Euclidean, TOMA Manhattan, and TOMA Chebyshev, as per the distance metric used in Eq. (2) (abbreviated as EUCL, MANH, and CHEB henceforth). We compare these to two state-of-the-art baselines, CAM [35] and MM [37].

Given a set of aspects A[8], CAM aggregates their scores through a weighted average:

$$\text{CAM}(r_t) = \sum_{a \in A} p_a \times \mu(\hat{r}_{t,a}) \tag{5}$$

where $\mu(\cdot)$ is the evaluation measure (e.g., NDCG), $\hat{r}_{t,a}$ is the ranking labelled with respect to aspect a, and p_a is a parameter controlling the importance of each aspect: $p_a \in [0,1]$ and $\sum_{a \in A} p_a = 1$.

MM [37] aggregates the evaluation measure scores computed for each aspect individually with a weighted harmonic mean:

$$\text{MM}(r_t) = \frac{\sum_{a \in A} p_a}{\sum_{a \in A} \frac{p_a}{\mu(\hat{r}_{t,a})}} \tag{6}$$

with the same notation as above.

Out of the other multi-aspect methods presented in §2, we do not use WHAM [35] as baseline because it also uses the weighted harmonic mean to aggregate the evaluation measure scores. However, WHAM is defined only for relevance and credibility, and can therefore be seen as an instantiation of MM restricted to two aspects. All other multi aspect measures in §2 need a predefined set and number of aspects, thus are not applicable in our scenario.

We instantiate our method and the baselines using (1) NDCG [31] and graded labels (when available); and using (2) AP [7] and binary labels (we convert all graded labels to binary by treating all grades above zero as one, and grades equal/below zero as zero). We consider all aspects equally important (all aspects are mapped to an integer scale with one unit separating each grade). All source code will be released upon publication of the paper.

4.2 Anomalies of CAM & MM

Next we discuss anomalies of CAM and MM that TOMA overcomes.

Problem 1: MM is ill-defined. As the harmonic mean is not defined with zero values, MM is not defined if $\exists a \in A$ such that $\mu(\hat{r}_{t,a}) = 0$, e.g., a ranking does not retrieve any correct or relevant document. To compute MM even in these cases, as the denominator in Eq. (6) tends to $+\infty$ if any $\mu(\hat{r}_{t,a})$ tends to zero, we set $\text{MM}(r_t) = 0$. For classification measures, this problem is called the Strong Definiteness Axiom [41]. It represents a serious issue for collections where there are a few documents with a positive label for certain aspects. For example, for the Task Tracks, since useful documents are very sparse, many systems are not able to retrieve any useful documents and they all have a 0 score, independently of the number of relevant documents they retrieve. TOMA does not have this problem, because we first assign a weight to each tuple of labels and then

[8] CAM was originally formulated for two aspects [35].

Table 3: CAM, MM and TOMA scores instantiated with AP & NDCG for all rankings in D. The highest scores are in bold.

AP

Length 3	CAM	MM	EUCL	MANH	CHEB	Length 2	CAM	MM	EUCL	MANH	CHEB	Length 1	CAM	MM	EUCL	MANH	CHEB
(d_1, d_2, d_3)	**0.7917**	**0.3684**	**1**	**1**	0.5	(d_1, d_2)	0.6250	0.25	**1**	**1**	0.5	(d_1)	0.5	0	0.5	0.5	0
(d_1, d_3, d_2)	**0.7917**	**0.3684**	0.8333	0.8333	0.3333	(d_1, d_3)	0.6250	0.25	0.5	0.5	0	(d_2)	0.25	0	0.5	0.5	**1**
(d_2, d_1, d_3)	0.6667	0.3125	**1**	**1**	**1**	(d_2, d_1)	0.5	0.25	**1**	**1**	**1**	(d_3)	0.25	0	0	0	0
(d_2, d_3, d_1)	0.6667	0.25	0.8333	0.8333	**1**	(d_2, d_3)	0.5	0	0.5	0.5	**1**	-	-	-	-	-	-
(d_3, d_1, d_2)	0.6667	0.3125	0.5833	0.5833	0.3333	(d_3, d_1)	0.5	0.25	0.25	0.25	0	-	-	-	-	-	-
(d_3, d_2, d_1)	0.6667	0.25	0.5833	0.5833	0.5	(d_3, d_2)	0.5	0	0.25	0.25	0.5	-	-	-	-	-	-

NDCG

Length 3	CAM	MM	EUCL	MANH	CHEB	Length 2	CAM	MM	EUCL	MANH	CHEB	Length 1	CAM	MM	EUCL	MANH	CHEB
(d_1, d_2, d_3)	**0.9073**	0.4489	0.9367	0.9711	0.8597	(d_1, d_2)	0.7682	0.3491	0.8080	0.8147	0.8597	(d_1)	0.4728	0.1491	0.4290	0.4693	0.3801
(d_1, d_3, d_2)	0.8824	0.4386	0.8917	0.9404	0.7602	(d_1, d_3)	0.6483	0.3145	0.5914	0.6667	0.3801	(d_2)	0.4682	0.2258	0.6006	0.5475	0.7602
(d_2, d_1, d_3)	0.9056	**0.4516**	**1**	**1**	**1**	(d_2, d_1)	0.7665	0.3776	0.8713	0.8436	**1**	(d_3)	0.2781	0	0.2574	0.3129	0
(d_2, d_3, d_1)	0.8801	0.4319	0.9775	0.9795	0.9502	(d_2, d_3)	0.6437	0.2679	0.7630	0.7449	0.7602	-	-	-	-	-	-
(d_3, d_1, d_2)	0.8106	0.3930	0.8284	0.8827	0.6199	(d_3, d_1)	0.5765	0.2801	0.5281	0.6089	0.2398	-	-	-	-	-	-
(d_3, d_2, d_1)	0.8100	0.3827	0.8509	0.8929	0.6697	(d_3, d_2)	0.5735	0.1897	0.6364	0.6583	0.4796	-	-	-	-	-	-

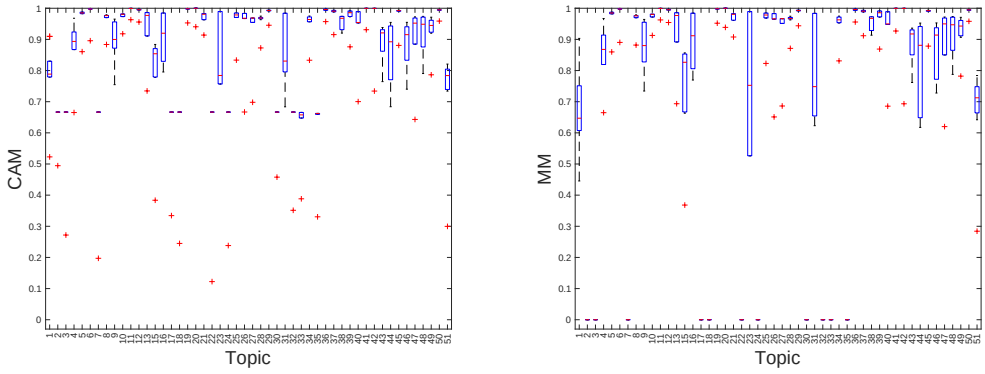


Figure 2: Box-plots for CAM and MM with NDCG on the Decision Track 2019. Topic numbers are on the x-axis and measures scores on the y-axis. The maximum achievable value for CAM and MM is variable and depends on the topic and the aspects.

compute a single-aspect evaluation measure μ, thus there is no division by 0 and TOMA is well defined.

Problem 2: CAM and MM can range in different intervals. Given a set of documents D and a set of aspects A, by definition CAM and MM are multi-aspect evaluation measures $M\colon D^* \to [0, X]$, where D^* is the set of rankings and $X \leq 1$. Depending on D and A, there exist cases with $X < 1$.

To prove this claim, we need to show that when M is CAM or MM, $\exists\ D, A$ such that:

$$\max_{r \in D^*} M(\hat{r}_t) < 1 \tag{7}$$

i.e., for each ranking of documents in D^* the maximum measure score will be less than 1. To build such an example, the set D needs to contain documents with not comparable tuples of labels:

$$\begin{aligned} &\exists d_1, d_2 \in D\colon \mathrm{GT}(d_1) \not\preceq \mathrm{GT}(d_2) \text{ and } \mathrm{GT}(d_2) \not\preceq \mathrm{GT}(d_1) \iff \\ &\exists d_1, d_2 \in D \text{ and } \exists a_1, a_2 \in A\colon \\ &\mathrm{GT}_{a_1}(d_1) \prec_{a_1} \mathrm{GT}_{a_1}(d_2) \text{ and } \mathrm{GT}_{a_2}(d_2) \prec_{a_2} \mathrm{GT}_{a_2}(d_1) \end{aligned} \tag{8}$$

In this case, CAM and MM cannot achieve a score equal to 1, as illustrated by the following example.

Consider the example in §3.4 with $A = \{\text{relevance, correctness}\}$. Let D be a set with 3 documents: $D = \{d_1, d_2, d_3\}$ that have the following labels:

$$\mathrm{GT}(d_1) = (\mathtt{mr}, \mathtt{c}) \quad \mathrm{GT}(d_2) = (\mathtt{hr}, \mathtt{pc}) \quad \mathrm{GT}(d_3) = (\mathtt{hr}, \mathtt{nc}) \tag{9}$$

Documents (d_1, d_2) and (d_1, d_3) are not comparable and there is no unequivocal way of sorting them, e.g., it is not clear if d_1 should be ranked before d_2 or vice-versa.

Let us consider CAM and MM instantiated with AP and NDCG. For AP we use a harsh mapping for relevance and correctness, i.e., $\{\mathtt{fr}, \mathtt{hr}\} \mapsto 1$ and $\{\mathtt{mr}, \mathtt{nr}\} \mapsto 0$, and $\mathtt{c} \mapsto 1$ and $\{\mathtt{pc}, \mathtt{nc}\} \mapsto 0$. For NDCG we map each category to a different integer, for relevance we have: $\mathtt{hr} \mapsto 15$, $\mathtt{fr} \mapsto 10$, $\mathtt{mr} \mapsto 5$, $\mathtt{nr} \mapsto 0$, and for correctness we have: $\mathtt{c} \mapsto 10$, $\mathtt{pc} \mapsto 5$, $\mathtt{nc} \mapsto 0$. The NDCG ideal ranking [29] for relevance is: $(\mathtt{hr}, \mathtt{hr}, \mathtt{mr})$ and for correctness is $(\mathtt{c}, \mathtt{pc}, \mathtt{nc})$. The NDCG log base is set to 2.

For TOMA we use the embedding of the first row in Tab. 1 and as weight function we map each equivalence class to a different

integer with step 1. We instantiate TOMA with AP and NDCG with log base 2 (Tab. 3). Since AP does not handle multi-graded weights, we map the top half of the equivalence classes to 1 and the rest to 0. Tab. 3 shows CAM, MM, and TOMA scores instantiated with AP NDCG for each possible ranking of documents in D.

In Tab. 3 none of the rankings in D^* can achieve a score equal to 1 for CAM and MM, while TOMA has at least one ranking with score 1. In CAM and MM this happens because, any way we sort the documents, either we penalize correctness, e.g., (d_2, d_3, d_1) or we penalize relevance, e.g., (d_1, d_2, d_3). TOMA does not have this problem, since it first defines how to sort tuples of labels, then weights them accordingly and computes the measure score. Thus, if we sort documents in the order induced by $\preceq_*$, we obtain a score equal to 1 (proof in the appendix). Experiments on real data confirm this, as detailed next.

Estimating CAM and MM Upper Bound. With the following experiment we show that with real data CAM and MM can be upper bounded by a value X lower than 1. To estimate the value X with real data, we generate different ideal rankings of documents with different strategies. The intuition is that by ranking documents in the best possible way, we should achieve a score equal to 1, as it happens for any single-aspect evaluation measure computed against the ideal ranking. Since CAM and MM do not define how to sort documents, i.e., a total order relation $\preceq_*$, we need to test different possible strategies to build these ideal rankings.

First, we define the ideal rankings obtained with a recursive strategy: these are the ideal rankings for each aspect when considered separately, e.g., for 3 aspects, a_1, a_2 and a_3, with a preference order where a_1 is followed by a_2, followed by a_3: (1) we sort the documents with decreasing label for a_1; (2) among the documents with the same label for a_1, we sort the documents with decreasing label for a_2; (3) among the documents with the same label for a_1 and a_2, we sort the documents with decreasing label for a_3. We generate these ideal rankings for each possible preference order among the aspects.

We also generate 3 additional ideal rankings: (1) we sum the weights across aspects and sort the documents by this sum; (2) we sum the squared weights across aspects and sort the documents by this sum; (3) we consider the highest weight across aspects and sort documents by their highest weight regardless of the aspect.

Fig. 2 reports the distributions of CAM with AP scores for the ideal rankings for the Decision Track 2019. These distributions depend on the aspect and the topic. We see that the upper bound X is variable and depends on the topic: just for 2% of topics it is equal to 1 and for 26% topics it is lower than 0.9. We obtain similar or even more extreme distributions of scores for all the other tracks (except for Misinformation 2020, see online appendix).

Interpretability of CAM and MM scores. Problem 2 is especially important because it affects the interpretability of CAM and MM scores. When a measure is used to assess the quality of a single ranking in isolation, it should be intuitively interpretable [33], e.g., NDCG=0.6 has the intuitive interpretation that the ranking can be further improved by 0.4. If TOMA is instantiated with NDCG, the intuitive interpretability of NDCG holds, but if CAM or MM are instantiated with NDCG, the intuitive interpretability of NDCG is lost: by the arguments above, CAM and MM may fail to obtain an optimal score of 1, and the optimal score depends on A and D, hence it cannot in general be known a priori.

This issue is important for MM, which is affected also by Problem 1, and therefore may have $X << 1$. Thus MM scores can be compressed towards 0, and this can lead to cases with many ties, where it is hard to distinguish between different rankings.

4.3 Experimental Findings

Empirically, evaluation measures are commonly assessed in terms of their correlation [24], discriminative power [39], informativeness [6], intuitiveness [40], and unanimity [2]. Out of these, we report only correlation and discriminative power because the rest does not apply: the informativeness test [6] requires a precision recall-curve, which cannot be defined for multi-aspect evaluation; the intuitiveness test [40] requires simple single-aspect measures (e.g. precision, recall), which do not apply to multi-aspect evaluation; the unanimity test [2], which is defined for multi-aspect evaluation, requires that all the simple measures agree over all aspects, which happened extremely rarely in our data, especially as the number of aspects increased (see the low correlation among aspects in Tab. 4).

4.3.1 Correlation Analysis. We use Kendall's τ [32] to estimate TOMA's correlation to MM and CAM. Generally, if a new evaluation measure *strongly* correlates to an existing one, it is likely to represent redundant information [46]. We use Kendall's τ because it has better gross-error sensitivity than the Pearson correlation coefficient [22], and because the Spearman correlation coefficient cannot handle ties. As per [24], we compute the correlation topic-by-topic. For each topic we consider the *Rankings of Submitted runs (RoS)* corresponding to two different measures (one ranking per measure) and then compute Kendall's τ between the two RoS. We report Kendall's τ averaged across all topics. As per [43, 44], we consider two rankings equivalent if Kendall's τ is greater than 0.9.

Tab. 4 shows the findings, which are summarised as follows:

- The RoS corresponding to EUCL - MANH are equivalent ($\tau = 1$) at all times for AP. This perfect correlation for AP happens because, by definition, when the sets of equivalence classes from these approaches are mapped to binary labels, they produce the exact same set of labels (see also Tab. 3). For NDCG, $\tau = 0.19 - 1$, where higher correlations correspond to tracks where some aspects are not assessed for non relevant documents, thus there are less extreme cases and EUCL is more similar to MANH.
- The RoS corresponding to (EUCL, MANH) - CHEB are very weakly correlated ($\tau = 0.01 - 0.32$), this is due to Chebyshev distance being very harsh, since many equivalence classes are considered equivalent to the class of non relevant documents.
- The RoS corresponding to EUCL - CAM and MANH - CAM are very weakly correlated ($\tau = 0.11 - 0.41$) for the Web tracks, but moderately correlated ($\tau = 0.54 - 0.76$) for the Task, Decision and Misinformation tracks. This happens because: (i) the runs submitted to the Web tracks were not designed to account for multiple aspects and (ii) for the Task, Decision and Misinformation tracks, usefulness, credibility and correctness are not assessed for non relevant documents.

Table 4: Kendall's τ correlation between rankings of systems and discriminative power (the higher, the better; best is in bold). Not all aspect combinations occur in all tracks (marked grey).

	WEB2009		WEB2010		WEB2011		WEB2012		WEB2013		WEB2014		TASK15		TASK16		DECISION19		MISINFO 2020	
	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP	NDCG	AP
	CORRELATION																			
EUCL - CAM	0.25	0.16	0.18	0.12	0.21	0.11	0.31	0.26	0.22	0.12	0.30	0.23	0.68	0.54	0.63	0.55	0.76	0.60	0.72	0.51
EUCL - MM	0.07	0.04	0.05	0.00	0.06	0.03	0.16	0.13	0.12	0.04	0.17	0.04	0.36	0.17	0.02	-0.10	0.46	0.31	0.45	0.27
MANH - CAM	0.22	0.16	0.21	0.12	0.16	0.11	0.34	0.26	0.31	0.12	0.41	0.23	0.62	0.54	0.60	0.55	0.69	0.60	0.72	0.51
MANH - MM	0.06	0.04	0.01	0.01	0.06	0.03	0.16	0.13	0.11	0.04	0.15	0.04	0.28	0.17	-0.06	-0.10	0.41	0.31	0.45	0.27
CHEB - CAM	0.01	0.02	0.06	0.05	0.02	0.02	0.19	0.15	0.11	0.00	0.19	0.07	0.26	0.16	-0.13	-0.18	0.27	0.27	0.29	0.24
CHEB - MM	0.06	0.09	0.00	0.03	0.09	0.01	0.19	0.16	0.12	0.08	0.14	0.05	0.88	0.86	0.52	0.53	0.58	0.60	0.54	0.52
EUCL - MANH	0.36	1.00	0.19	1.00	0.21	1.00	0.34	1.00	0.20	1.00	0.34	1.00	0.87	1.00	0.71	1.00	0.72	1.00	1.00	1.00
EUCL - CHEB	0.01	0.02	0.10	0.03	0.01	0.03	0.32	0.11	0.22	0.04	0.30	0.12	0.33	0.19	-0.21	-0.21	0.28	0.21	0.26	0.20
MANH - CHEB	0.01	0.02	0.03	0.03	0.02	0.03	0.21	0.11	0.10	0.04	0.18	0.12	0.32	0.19	-0.24	-0.21	0.25	0.21	0.26	0.20
CAM - MM	0.10	0.05	0.05	0.01	0.11	-0.01	0.23	0.13	0.16	0.03	0.26	0.04	0.30	0.18	0.09	0.00	0.51	0.41	0.51	0.42
	CORRELATION																			
Relevance - Popularity	0.03	0.05	0.01	0.0	0.01	0.02	0.09	0.09	0.06	0.01	0.07	0.02	0.04	0.04	-0.03	0.01				
Relevance - Non-spam	0.05	0.03	0.02	0.0	0.03	0.01	0.07	0.05	-0.02	-0.01	0.07	-0.01	0.25	0.17	-0.07	-0.08				
Popularity - Non-spam	0.04	0.03	0.02	-0.01	0.04	0.01	0.07	0.04	-0.03	-0.02	0.04	0.00	0.07	0.02	0.08	0.06				
Relevance - Usefulness													0.75	0.75	0.75	0.74				
Usefulness - Popularity													0.10	0.06	-0.04	0.00				
Usefulness - Non-spam													0.40	0.33	-0.16	-0.19				
Credibility - Correctness																	0.26	0.26	0.28	0.24
Relevance - Credibility																	0.33	0.33	0.29	0.25
Relevance - Correctness																	0.42	0.49	0.49	0.47
	DISCRIMINATIVE POWER OF MEASURES																			
CAM	75.98	64.43	66.32	61.23	75.14	61.64	**68.71**	56.74	76.89	57.05	85.06	78.85	53.33	33.33	72.22	55.56	72.58	70.56	71.53	70.90
MM	75.61	50.58	**72.89**	**67.79**	67.32	67.81	62.68	56.12	**80.71**	46.99	74.25	53.56	0.00	0.00	0.00	0.00	60.08	53.23	68.31	62.20
EUCL	75.29	72.64	62.96	66.75	75.14	70.33	66.13	64.10	75.14	**59.45**	80.92	78.85	**66.67**	**66.67**	69.44	**75.00**	73.59	**73.99**	72.86	**75.14**
MANH	**76.66**	**72.68**	63.59	67.14	**77.32**	**70.38**	66.05	**64.18**	76.67	59.34	**86.44**	**79.08**	**66.67**	53.33	**75.00**	**75.00**	**73.79**	73.79	**73.02**	74.98
CHEB	50.18	6.32	59.82	51.49	73.06	50.11	61.08	39.36	77.10	49.34	75.17	66.21	0.00	0.00	0.00	0.00	42.54	29.84	65.41	53.33

Therefore, since some of the values are missing, these methods generate a lower number of equivalence classes, which make them more similar to CAM. Whereas, for the Web tracks, popularity and non-spamminess are approximated for all documents, meaning that MANH and EUCL can possibly generate all the different equivalence classes, even for non relevant documents. This makes them less similar to CAM than on the Task or Decision tracks.

- For the Task, Decision and Misinformation tracks, the RoS corresponding to MM and CHEB are moderately correlated ($\tau = 0.52 - 0.88$). The fact that, for these tracks, usefulness, credibility and correctness are not assessed for non relevant documents, means that all the documents that are mapped to a 0 weight with CHEB, are also contributing as 0 to MM.

To contextualise these findings, the middle part of Tab. 4 shows the τ values of the RoS corresponding to evaluating a single aspect only. Overall, the resulting correlations are low to non-existent, meaning that considering multiple aspects affects the final evaluation outcome. The two exceptions where the correlation between RoS is not very low are:

- For Task 2015-2016, for relevance - usefulness, $\tau = 0.74 - 0.75$. This happens because: (1) usefulness is not assessed for non relevant documents, thus non relevant documents are assumed to be not useful, and (2) usefulness is a very sparse signal (1.75% of documents are useful).
- For the Decision and Misinformation Tracks, for all aspects, $\tau = 0.24 - 0.49$. Again here credibility and correctness are not assessed for non relevant documents (6.89% of documents are credible and 9.75% are correct for the Decision Track; 13.73% of documents are correct and 27.62% are credible for the Misinformation Track), so the correlation is not as high as for the Task tracks.

Overall, the most correlated RoS correspond to: EUCL - MANH (τ up to 1), (EUCL, MANH)- CAM (τ up to 0.76), and CHEB - MM (τ up to 0.88). Intuitively, EUCL and MANH may be more similar to CAM (mean), while CHEB may be more similar to MM (harmonic mean). Thus TOMA proposes an alternative evaluation framework, which overcomes CAM and MM anomalies (see §4.2). The fact that τ values between TOMA and the baselines are never above 0.9 means that there are noticeable differences between the RoS generated by TOMA and by CAM or MM. Recall that all measures are instantiated with NDCG or AP, meaning that differences between them are due to how multi-aspect labels are treated.

4.3.2 Discriminative Power. We use Bootstrap Hypothesis Test [39] to estimate the discriminative power of TOMA, CAM and MM. Given a set of topics and a set of runs, we first generate subsets of topics by sampling with replacement the complete set of topics. We set the number of bootstrap samples to 10 000. To assess whether the measure scores for pairs of runs can be considered different at a given confidence level, we use a Paired Bootstrap Hypothesis Test. The confidence level is $1 - \alpha$, where α is the Type I Error, i.e., the probability to consider two systems different even if they are equivalent. We set $\alpha = 0.01$, requiring strong evidence for two systems to be different.

Tab. 4 (bottom part) displays the results of the discriminative power analysis, where the higher the score, the more discriminative (i.e., the better) the approach. We see that 16/20 times either MANH (12/20) or EUCL (6/20)[9] is best. The remaining 4 times, MM is best

[9]Ties are included in these counts.

Table 5: Number of times (%) that the labels of all aspects sum to 0 for a document that is ranked at position 1-5 (column 1) in a run that has been assessed as best per {topic, track, year} separately with {CAM, MM, EUCL, MANH, CHEB} using a retrieval cutoff of 5. The lower, the better.

Rank	CAM	MM	EUCL	MANH	CHEB
1	51 (1.18%)	131 (3.02%)	39 (0.90%)	**33 (0.76%)**	154 (3.55%)
2	65 (1.50%)	159 (3.67%)	50 (1.15%)	**48 (1.11%)**	179 (4.13%)
3	103 (2.38%)	202 (4.66%)	88 (2.03%)	**78 (1.80%)**	185 (4.17%)
4	102 (2.35%)	173 (3.99%)	86 (1.99%)	**74 (1.71%)**	183 (4.23%)
5	107 (2.47%)	196 (4.53%)	95 (2.19%)	**81 (1.87%)**	205 (4.73%)
1-5	428 (9.88%)	861 (19.88%)	358 (8.27%)	**314 (7.25%)**	906 (20.92%)

Table 6: Average sum of aspect labels for a document that is ranked at position 1-100 (column 1) in a run that has been assessed as best per {topic, track, year} separately with {CAM, MM, EUCL, MANH, CHEB} using a retrieval cutoff of 100. The higher, the better.

Ranks	CAM	MM	EUCL	MANH	CHEB
1-25	**1.70**	1.49	1.67	1.69	1.39
26-50	0.85	0.78	0.91	**0.94**	0.70
51-75	0.57	0.53	0.63	**0.64**	0.48
76-100	0.40	0.39	0.43	**0.44**	0.36

3 times, and CAM once. We also see that CHEB is never best, and for Task 2015-2016 it is actually zero. This is due to the very small amount of positive labels for usefulness in that track. For the same reason, MM is also zero for the same track. Overall, CHEB is the least discriminative measure, followed by MM; this is due to how these methods treat tuples of labels: the fact that if one aspect label is zero, then the whole score is zero, practically means that many runs are considered equal purely on that basis.

4.3.3 Zero-aspect documents. Our next analysis is motivated by the empirical trash@k measure often used in industry to mitigate the high cost of retrieving "trash" in high ranks. We count how often the labels of all aspects sum to zero for a document that has been ranked at position 1-5 in a run that has been assessed as the best run per track year, on a per query basis, using a retrieval cutoff of 5, separately with {CAM, MM, EUCL, MANH, CHEB} when instantiated separately with NDCG and AP. When the labels of all aspects sum to zero, this means that the corresponding document is of the worst quality. Ideally, such documents should not be retrieved, but when they do, they should not be in the top 5.

In Tab. 5 we see that MANH is associated with the lowest amount of zero-aspect documents, closely followed by EUCL. This happens because MANH is designed so that the higher the sum of a document's labels across aspects, the better that document will be deemed. CHEB is overall worst, closely followed by MM. This closeness between EUCL-MANH and CHEB-MM agrees with the previous correlation and discriminative power analysis. Overall, MANH (and less so EUCL) penalise low quality documents the best.

4.3.4 Document quality @1-100. We look at the quality of documents that have been ranked at positions 1-100 in a run that has been assessed as best per {topic, track, year} separately with {CAM, MM, EUCL, MANH, CHEB}, when instantiated separately with NDCG and AP, using a retrieval cutoff of 100. We split the ranks 1-100 into four sets (1-25, 26-50, 51-75, 76-100). For each document in each set, we sum the labels of its aspects, and we report the average of these sums, which can be seen as an approximation of the average document quality (the higher, the better).

As expected, we see that the numbers in Tab. 6, and hence document quality, drop as we move down the ranking, at all times. Comparing across columns however, we see that, for the runs that were assessed as best by MANH, document quality is overall, albeit marginally, the best, at ranks 26-100. This illustrates that the design of MANH (the higher the sum of a document's labels across aspects, the better that document will be considered) gives it the practical advantage of, not only reducing the amount of low quality documents in the top ranks (as seen in Tab. 5), but also of increasing the quality of documents further down the ranking, as we see now. Again, as previously, we observe that EUCL is a close second-best method, CHEB and MM are overall worst, and CAM is in between (although best, together with MANH, for the top ranks).

5 CONCLUSION AND LIMITATIONS

Multi-aspect evaluation is a special case of IR evaluation where the ranked list of documents returned by an IR system in response to a query must be assessed in terms of not only relevance to the query, but also other *aspects* (or dimensions) of the ranked documents, e.g., credibility or usefulness. We presented a principled multi-aspect evaluation approach, called TOMA, that is defined for any number and type of aspect, and that allows for (i) aspects having different gradings, (ii) any relative importance weighting for different aspects, and (iii) integration with any existing single-aspect evaluation measure, such as NDCG. We showed that TOMA has better discriminative power than prior approaches to multi-aspect evaluation, and that it is better at rewarding high quality documents across the ranking.

One limitation of TOMA is represented by the arbitrary choices in its definition: the embedding function, the distance function and the weight function. The embedding function maps labels from a nominal or ordinal scale to an interval or ratio scale. This calls for a in-depth investigation of the theoretical properties of TOMA using the existing axiomatic treatments of effectiveness for IR retrieval measures [3, 4, 9, 23, 36]. This also motivates a deep analysis of the interactions between different aspects and/or documents and how to handle them with TOMA, for example by defining a proper representation and distance function in a vector space which accounts for aspects as diversity, novelty, and redundancy. Moreover, the embedding function combined with the distance function can generate a large number of tuple of labels, which can be mapped to different integers through the weight function. This might represent a problem for gain based measures, thus a possible solution is to use TOMA to define the ideal ranking and then use effectiveness measures based on similarity to ideal rankings [17–19]. Finally, the empirical impact of varying both distance functions and weight functions should also be investigated, as should the impact of employing further multi-graded measures such as *Expected Reciprocal Rank (ERR)* [10], and the alignment of our current approach with real user preferences.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. 2009. Diversifying Search Results. In *Proc. 2nd ACM International Conference on Web Searching and Data Mining (WSDM 2009)*, R. Baeza-Yates, P. Boldi, B. Ribeiro-Neto, and B. B. Cambazoglu (Eds.). ACM Press, New York, USA, 5–14.

[2] A. Albahem, , D. Spina, F. Scholer, and L. Cavedon. 2019. Meta-evaluation of Dynamic Search: How Do Metrics Capture Topical Relevance, Diversity and User Effort?. In *Advances in Information Retrieval. Proc. 41st European Conference on IR Research (ECIR 2019)*, L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra (Eds.). Lecture Notes in Computer Science (LNCS) 10772, Springer, Heidelberg, Germany, 607–620.

[3] E. Amigó, F. Giner, S. Mizzaro, and D. Spina. 2018. A Formal Account of Effectiveness Evaluation and Ranking Fusion. In *Proc. 4th ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2018)*, D. Song, T.-Y. Liu, L. Sun, P. Bruza, M. Melucci, F. Sebastiani, and G. Hui Yang (Eds.). ACM Press, New York, USA, 123–130.

[4] E. Amigó and S. Mizzaro. 2020. On the Nature of Information Access Evaluation Metrics: a Unifying Framework. *Information Retrieval Journal* 23, 3 (2020), 318–386. https://doi.org/10.1007/s10791-020-09374-0

[5] E. Amigó, D. Spina, and J. Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *Proc. 41th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*, K. Collins-Thompson, Q. Mei, B. Davison, Y. Liu, and E. Yilmaz (Eds.). ACM Press, New York, USA, 625–634.

[6] J. A. Aslam, E. Yilmaz, and V. Pavlu. 2005. The maximum entropy method for analyzing retrieval measures. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, Ricardo A. Baeza-Yates, Nivio Ziviani, Gary Marchionini, Alistair Moffat, and John Tait (Eds.). ACM, 27–34. https://doi.org/10.1145/1076034.1076042

[7] C. Buckley and E. M. Voorhees. 2005. Retrieval System Evaluation. In *TREC. Experiment and Evaluation in Information Retrieval*, D. K. Harman and E. M. Voorhees (Eds.). MIT Press, Cambridge (MA), USA, 53–78.

[8] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. 2005. Learning to Rank using Gradient Descent. In *Proc. 22nd International Conference on Machine Learning (ICML 2005)*, S. Dzeroski, L. De Raedt, and S. Wrobel (Eds.). ACM Press, New York, USA, 89–96.

[9] L. Busin and S. Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proc. 4th International Conference on the Theory of Information Retrieval (ICTIR 2013)*, O. Kurland, D. Metzler, C. Lioma, B. Larsen, and P. Ingwersen (Eds.). ACM Press, New York, USA, 22–29.

[10] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009)*, D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin (Eds.). ACM Press, New York, USA, 621–630.

[11] C. Clarke, N. Craswell, and I. Soboroff. 2009. Overview of the TREC 2009 Web Track. In *TREC*. National Institute of Standards and Technology (NIST).

[12] Charles Clarke, Maria Maistro, Mark Smucker, and Guido Zuccon. 2020. Overview of the TREC 2020 Health Misinformation Track (to appear). In *Proceedings of the Twenty-Nine Text REtrieval Conference, TREC 2020, Gaithersburg, Maryland, USA, November 16-19, 2020 (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. -. National Institute of Standards and Technology (NIST).

[13] C. L. A. Clarke and G. V. Cormack. 2011. Overview of the TREC 2010 Web Track. In *TREC*. National Institute of Standards and Technology (NIST).

[14] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. 2011. Overview of the TREC 2011 Web Track. In *TREC*. National Institute of Standards and Technology (NIST).

[15] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *TREC*. National Institute of Standards and Technology (NIST).

[16] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani (Eds.). ACM Press, New York, USA, 659–666.

[17] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. 2020. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 225–234.

[18] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020. Offline Evaluation without Gain. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang, and Klaus Berberich (Eds.). ACM, 185–192.

[19] C. L. A. Clarke, A. Vtyurina, and M. D. Smucker. 2021. Assessing Top-k Preferences. arXiv:cs.IR/2007.11682

[20] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. 2014. TREC 2014 Web Track Overview. In *The Twenty-Second Text REtrieval Conference Proceedings (TREC 2013)*, E. M. Voorhees (Ed.). National Institute of Standards and Technology (NIST), Special Publication 500-302, Washington, USA.

[21] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. 2015. TREC 2014 Web Track Overview. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-308, Washington, USA.

[22] C. Croux and C. Dehon. 2010. Influence Functions of the Spearman and Kendall Correlation Measures. *Statistical Methods & Applications* 19 (2010), 497–515. https://doi.org/10.1007/s10260-010-0142-z

[23] M. Ferrante, N. Ferro, and N Fuhr. 2021. Towards Meaningful Statements in IR Evaluation. Mapping Evaluation Measures to Interval Scales. *CoRR* abs/2101.02668 (2021). arXiv:2101.02668 https://arxiv.org/abs/2101.02668

[24] M. Ferrante, N. Ferro, and E. Losiouk. 2019. How do Interval Scales Help us with Better Understanding IR Evaluation Measures? *Information Retrieval Journal* (2019). https://doi.org/10.1007/s10791-019-09362-z

[25] M. Ferrante, N. Ferro, and M. Maistro. 2015. Towards a Formal Framework for Utility-Oriented Measurements of Retrieval Effectiveness. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2808194.2809452

[26] M. Ferrante, N. Ferro, and Silvia P. 2017. Are IR Evaluation Measures on an Interval Scale?. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 67–74. https://doi.org/10.1145/3121050.3121058

[27] M. Ferrante, N. Ferro, and S. Pontarollo. 2019. A General Theory of IR Evaluation Measures. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31, 3 (2019), 409–422.

[28] P. R. Halmos. 1974. *Naive Set Theory*. Springer-Verlag, New York, USA.

[29] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[30] G. Kazai, S. Masood, and M. Lalmas. 2004. A Study of the Assessment of Relevance for the INEX'02 Test Collection. In *Advances in Information Retrieval. Proc. 26th European Conference on IR Research (ECIR 2004)*, S. McDonald and J. Tait (Eds.). Lecture Notes in Computer Science (LNCS) 2997, Springer, Heidelberg, Germany, 296–310.

[31] J. Kekäläinen and K. Järvelin. 2002. Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science and Technology (JASIST)* 53, 13 (November 2002), 1120—1129.

[32] M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (November 1945), 239–251.

[33] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti (Eds.). ACM, 571–580. https://doi.org/10.1145/1772690.1772749

[34] C. Lioma, M. Maistro, M. D. Smucker, , and D. Zuccon. 2019. Overview of the TREC 2019 Decision Track. In *The Twenty-Eighth Text REtrieval Conference Proceedings (TREC 2019)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-331, Washington, USA.

[35] C. Lioma, J. G. Simonsen, and B. Larsen. 2017. Evaluation Measures for Relevance and Credibility in Ranked Lists. In *Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017)*, J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz (Eds.). ACM Press, New York, USA, 91–98.

[36] E. Maddalena and S. Mizzaro. 2014. Axiometrics: Axioms of Information Retrieval Effectiveness Metrics. In *Proc. 6th International Workshop on Evaluating Information Access (EVIA 2014)*, S. Mizzaro and R. Song (Eds.). National Institute of Informatics, Tokyo, Japan, 17–24.

[37] J. Palotti, G. Zuccon, and A. Hanbury. 2018. MM: A New Framework for Multidimensional Evaluation of Search Engines. In *Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018)*, A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Broder, M. J. Zaki, S. Candan, A. Labrinidis, A. Schuster, and H. Wang (Eds.). ACM Press, New York, USA, 1699–1702.

[38] C. Petersen, C. Lioma, J. G. Simonsen, and B. Larsen. 2015. Entropy and Graph Based Modelling of Document Coherence Using Discourse Entities: An Application to IR. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. ACM, New York, NY, USA, 191–200. https://doi.org/10.1145/2808194.2809458

[39] T. Sakai. 2007. Evaluating Information Retrieval Metrics Based on Bootstrap Hypothesis Tests. *IPSJ Digital Courier* 3 (2007), 625–642. https://doi.org/10.2197/ipsjdc.3.625

[40] T. Sakai. 2012. Evaluation with Informational and Navigational Intents. In *Proc. 21st International Conference on World Wide Web (WWW 2012)*, A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab (Eds.). ACM Press, New York, USA, 499–508.

[41] F. Sebastiani. 2015. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. In *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, and Y. Zhang (Eds.). ACM Press, New York, USA, 11–20.

[42] Z. Tang and G. H. Yang. 2017. Investigating per Topic Upper Bound for Session Search Evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 185–192. https://doi.org/10.1145/3121050.3121069

[43] E. M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel (Eds.). ACM Press, New York, USA, 315–323.

[44] E. M. Voorhees. 2001. Evaluation by Highly Relevant Documents. In *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel (Eds.). ACM Press, New York, USA, 74–82.

[45] E. M. Voorhees and A. Ellis (Eds.). 2016. *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*. Vol. Special Publication 500-321. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec25/trec2016.html

[46] W. Webber, A. Moffat, J. Zobel, and T. Sakai. 2008. Precision-at-ten Considered Redundant. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 695–696. https://doi.org/10.1145/1390334.1390456

[47] Y. Y. Yao. 1995. Measuring Retrieval Effectiveness Based on User Preference of Documents. *Journal of the American Society for Information Science (JASIS)* 46, 2 (1995), 133–145. https://doi.org/10.1002/(SICI)1097-4571(199503)46:2<133::AID-ASI6>3.0.CO;2-Z

[48] E. Yilmaz, M. Verma, R. Mehrotra, E. Kanoulas, B. Carterette, and N. Craswell. 2016. Overview of the TREC 2015 Tasks Track. In *The Twenty-Fourth Text REtrieval Conference Proceedings (TREC 2015)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-319, Washington, USA.

[49] G. Zuccon. 2016. Understandability Biased Evaluation for Information Retrieval. In *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*, N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello (Eds.). Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany, 280–292.

10. *Test collection incompleteness and unjudged documents*

Lucas Chaves Lima, Casper Hansen, Christian Hansen, Maria Maistro, Jakob Grue Simonsen, Christian Lioma. Test collection incompleteness and unjudged documents (Under Revision). 2021, 30th ACM International Conference on Information and Knowledge Management (CIKM2021) [93]

Test collection incompleteness and unjudged documents

Anonymous Author(s)

ABSTRACT

IR test collections are notoriously *incomplete*: they contain considerably fewer assessed than non-assessed documents, and within this minority class of assessed documents, considerably fewer documents are assessed as relevant than non-relevant to a query. One state-of-the-art way of addressing this problem is to automatically infer relevance assessments from document similarities. However, this is currently done without accounting for the imbalance between relevant and non-relevant documents, which means that the discriminative signal of relevant documents (minority) is weakened by the much stronger noise signal of non-relevant documents (majority). We address this with a simple method of reducing the noisy impact of highly dissimilar documents when inferring relevance assessments from document similarities. We show that our method is effective for inferring relevance assessments of non-assessed documents by performing experiments with 8 different sampling approaches, 3 different representations of document semantics (TF-IDF, Glove and BERT word embeddings), and data from 5 different TREC tracks (248 topics and 256 runs).

ACM Reference Format:
Anonymous Author(s). 2020. Test collection incompleteness and unjudged documents. In *CIKM '21: 30th ACM International Conference on Information and Knowledge Management, November 1-5 2021, Queensland, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

IR test collections are notoriously *incomplete*: not all documents are assessed, and many more documents are assessed as non-relevant than as relevant. This problem [3] has been generally addressed in three ways. One is to consider non-assessed documents as non-relevant, but this creates bias in favour of methods that contribute more to the pooling [3]. Another way is to ignore non-assessed documents altogether, but this can lead to unfair inferences, such as treating the ranking of non-assessed documents in the top k as equally good to the ranking of only relevant documents in the top k [16]. A third way is to reduce non-assessed documents by inferring relevance judgements automatically [1, 2, 4, 17, 19, 20]. Even if inferred judgments may be incorrect, having better-than-random judgments may (i) yield a fairer comparison between systems, and (ii) enable supervised or semi-supervised learning to be trained on larger datasets of annotated documents. There are different ways of automatically inferring judgements. Document relevance can be approximated from IR system rankings [1, 17, 19, 20]: if a document is ranked close to the top of the ranking by several different systems, then that document is considered relevant. Alternatively, we can infer the precision/recall curve first, and then use it to estimate the labels of non-assessed documents [2]. Both these approaches have limitations (see § 2). Another alternative is to infer the assessment of a document based on its similarity to a set of already assessed documents [4]. A problem with this approach is that it does not account for the imbalance between relevant and non-relevant documents: i.e. since the number of non-relevant documents is considerably higher than the number of relevant ones, the discriminative signal of relevant documents (minority) is weakened by the much stronger noise signal of non-relevant documents (majority). We address this, by **contributing** a tunable Document Similarity Threshold (DST) that reduces the noisy impact of the many highly dissimilar (non-relevant) documents.

CIKM '21, November 1-5, 2021, Queensland, Australia

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 RELATED WORK

Roitero et al. [15] survey unsupervised ways of inferring document labels and propose a method for combining them. None of these methods use topic/document content, but instead: if a document is placed close to the top of the ranking by many different systems, then that document is considered relevant [15, 17, 19, 20]. One of the main problems is that those approaches, stretched over time, can really penalize diversity of thought in terms of what is relevant.

Aslam and Yilmaz [2] infer the precision/recall curve first, and then use it to estimate the labels of unjudged documents. Since different distributions of relevant and non-relevant documents can possibly generate the same precision/recall curve, the labels estimated with this approach do not necessarily match the actual relevance of unjudged documents, so, this approach is not comparable to estimating the relevance of a specific document.

We relate to [4] instead, which is based on the *cluster hypothesis*: "closely associated documents tend to be relevant to the same request" [8, 22]. The relevance label of an unjudged document is inferred from the similarity between the unjudged document and a set of judged documents. However this does not account for the imbalance between the number of relevant and non-relevant documents: since the number of non-relevant documents is considerably higher than the number of relevant ones, the discriminative signal of relevant documents is weakened by the much stronger noise signal of non relevant documents. We overcome this with a simple learnable threshold θ described in § 3 combined with oversampling and undersampling approaches[1] described next.

A popular oversampling method is Synthetic Minority Oversampling Technique (SMOTE) [5]. In the feature representation space, SMOTE randomly selects similar samples from the minority class, draws a line between the sample representations and generates a synthetic sample as a point along that line. A crucial step in SMOTE is how to identify similar samples to be used as candidates for the generation of synthetic samples [10, 13]. KmeansSmote [10] uses

[1]Sampling methods aim to modify imbalanced datasets by expanding the minority class (oversampling) or reducing the majority class (undersampling).

a clustering approach (k-means) [12] to identify similar samples in the minority class and SMOTE to generate synthetic samples. Borderline Oversampling (BO) [13] trains a Support-Vector Machine (SVM) classifier to learn the decision boundary between the classes, and then uses SMOTE to generate synthetic samples of the minority class near the decision boundary. The Adaptive Synthetic sampling approach (ADASYN) [7], for each sample of the minority class, finds the k-Nearest Neighbours and calculates the ratio between the number of samples from the neighbourhood in the majority class and k. The higher this ratio, the more samples from the majority class are included in the neighbourhood. Thus this is a region where minority samples are "isolated", where it is beneficial to create synthetic samples. The new synthetic samples are then created by generating synthetic samples in between samples of the minority class.

Undersampling can be done by removing Tomek's links [21]. A Tomek's link is defined when two training samples are nearest neighbors, but belong to different classes. The assumption is that only noisy/boundary samples would have Tomek's links. Thus majority class samples that are Tomek-linked to minority class samples are removed. Clustering-Based Undersampling (CBU) [11] builds k clusters on the majority class, where k is equal to the number of samples in the minority class, and then uses the cluster centroids as the new synthetic samples of the majority class. Instance Hardness (IH) [18] predicts the instance hardness as the probability of each sample being misclassified by different classification algorithms. If a sample is often misclassified, it is considered a hard instance to predict, thus it is removed from the training set.

3 APPROACH

We present the method by Carterette and Allan [4] to infer relevance labels, and then how we address its limitation wrt. class imbalance with our Document Similarity Thresholding (DST) method.

Inferring Relevance Assessments [4]. Let $\mathcal{D}$ be a set of documents. For a topic t, $\mathcal{D} = \mathcal{A} \cup \bar{\mathcal{A}}$, where $\mathcal{A}$ and $\bar{\mathcal{A}}$ are the sets of *assessed*, resp. *non-assessed* documents wrt. t. Similarly, $\mathcal{A} = \mathcal{R} \cup \bar{\mathcal{R}}$, where $\mathcal{R}$ and $\bar{\mathcal{R}}$ are the sets of *relevant* and *non-relevant* documents wrt. t. Given a non-assessed document $\bar{d} \in \bar{\mathcal{A}}$, the goal is to predict the relevance assessment of $\bar{d}$ by exploiting its semantic similarity to the set of assessed documents $\mathcal{A}$. To do so, the probability of relevance of $\bar{d}$ is conditioned to its similarity to the documents in $\mathcal{A}$. Then, a logistic regression model learns the weight β, where β_d denotes the contribution of each document $d \in \mathcal{A}$ to the estimation of the relevance label for $\bar{d}$. The logistic regression model maximizes the following log-likelihood:

$$\log(\mathcal{L}(\beta)) = \sum_{d \in \mathcal{A}} (y_d \log(p_d) + (1 - y_d) \log(1 - p_d)) + \lambda \sum_{d \in \mathcal{A}} \beta_d^2 \quad (1)$$

where λ is a penalization parameter to avoid overfitting, p_d is the probability of d being relevant, $y_d = 1$ if $d \in \mathcal{R}$, $y_d = 0$ if $d \in \bar{\mathcal{R}}$. The log-odds of the probability of a non-assessed document $\bar{d}$ to be relevant are defined as a weighted sum:

$$\log\left(\frac{p_{\bar{d}}}{1 - p_{\bar{d}}}\right) = \beta_0 + \sum_{d \in \mathcal{A}} \beta_d \text{sim}(\bar{d}, d) \quad (2)$$

sim is a similarity function (we use cosine similarity as per [4]). The model returns $p_{\bar{d}}$, the probability of $\bar{d}$ being relevant, which we map to the relevance label with the highest probability.

Document Similarity Thresholding (DST). A problem with the above is that the imbalance of non-relevant documents (majority) and relevant documents (minority) affects directly the estimation of the probability $p_{\bar{d}}$ in Eq. (2). Let us assume that the correct label of the currently non-assessed document $\bar{d}$ is that it is relevant. When computing the similarity of $\bar{d}$ to the set of assessed documents, we argue that $\bar{d}$ should be more similar to documents that are assessed relevant to the topic, and less similar to documents that are assessed non-relevant to the topic. However, if the number of highly dissimilar documents to $\bar{d}$ in the set is very large, then *even if the individual contribution of each highly dissimilar document to Eq.* (2) *is low*, due to their numerical superiority, these dissimilar documents can possibly surpass the contribution of highly similar documents.

To reduce this effect, we define a decision boundary to filter out noisy documents. This decision boundary can be applied to any inferring label approach that relies on inter-document similarities. Specifically, we define a learnable threshold θ to filter out documents with low similarity to $\bar{d}$, hence reducing the noisy impact of highly dissimilar documents. Given a threshold θ, we define $A(\bar{d}, \theta)$ to be the set of assessed documents that have a similarity score $> \theta$ with respect to $\bar{d}$, that is: $A(\bar{d}, \theta) = \{d \in \mathcal{A} : \text{sim}(\bar{d}, d) > \theta\}$. Note that θ does not depend on $\bar{d}$, but it is rather a global parameter which depends on the topic. This step is applied to Eq. (2) as follows:

$$\log\left(\frac{p_{\bar{d}}}{1 - p_{\bar{d}}}\right) = \beta_0 + \sum_{d \in A(\bar{d}, \theta)} \beta_d \text{sim}(\bar{d}, d) \quad (3)$$

where $\mathcal{A}$ is replaced with $A(\bar{d}, \theta)$. This allows to discard a possibly high number of very dissimilar documents when estimating the probability of a document being relevant. Note that, if we do not use the threshold θ, even if a relevant document is very similar to $\bar{d}$, the majority of dissimilar documents will likely mitigate this signal.

Sampling Approaches to Infer Labels. DST can be combined with any sampling approach by modifying the representation of documents in the feature space. Formally, a sampling approach S maps a given set of assessed documents $\mathcal{A}$ to a new set of documents $S(\mathcal{A}) \mapsto \mathcal{A}^*$, where $\mathcal{A}^*$ is balanced, i.e. $|\mathcal{R}|/|\bar{\mathcal{R}}| \sim 1$. To apply any sampling method, first we need to define the feature representation of each document. We use the set of assessed documents $\mathcal{A}$ as reference and represent each document with the vector of its similarity scores wrt. the assessed documents: $d \in \mathcal{D}$ is mapped to its feature vector $\vec{d} = (sim(d, d_1), \ldots, sim(d, d_j))$ for all $j \in \{1, \ldots, |\mathcal{A}|\}$. Given the document representations in this feature space, the idea is to account for the threshold θ when representing documents. The new document representation is denoted by $\vec{d}(\theta) = (sim(d, d_1, \theta), \ldots, sim(d, d_j, \theta))$ for all $j \in \{1, \ldots, |\mathcal{A}|\}$, where

$$\text{sim}(d, d_j, \theta) = \begin{cases} \text{sim}(d, d_j) & \text{if sim}(d, d_j) > \theta \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

i.e. we only include documents that have similarity score $> \theta$.

4 EXPERIMENTAL EVALUATION

We experimentally evaluate our method by considering the task of inferring the relevance labels of unjudged documents. We compare: logistic regression to infer relevance labels [4] (baseline), combined with DST, the sampling approaches in § 2, and both DST and sampling approaches[2]. We use the TREC data shown below:

	Web 2010	Web 2011	Web 2012	Web 2013	Web 2014
Dataset	ClueWeb09	ClueWeb09	ClueWeb09	ClueWeb12	ClueWeb12
Topics	48	50	50	50	50
Submitted runs	56	61	48	61	30
Relevance grades	5	4	5	5	5
Percentage of Samples per Class after Mapping					
Not Relevant	79.19%	86.73%	76.78%	71.27%	60.83%
Partially Relevant	15.68%	10.28%	14.81%	20.96%	26.48%
Relevant	5.24%	6.71%	10.22%	8.71%	12.95%

While [4] used binary relevance labels, we use graded relevance labels. We merge negative (spam) and zero labels (non-relevant) into the non-relevant class, and all labels higher or equal than 2 (highly relevant and key) into the highly relevant class. When inferring relevance assessments, we represent documents using TF-IDF (as per [4]), average Glove word embedding [14][3], and the contextual BERT embedding [6][4]. As BERT is limited to 512 tokens, we split the text into 512-token segments and compute the average BERT embedding over all segments.

We refer to the logistic regression method [4] as LogReg (baseline) and as LogReg-DST when combined with DST. In addition to the sampling approaches from § 2, we use Random Oversampling (RO). In the online appendix we also report undersampling approaches [5]. When combining DST with a sampling approach, we first use DST to define the feature space as in Eq. (4), and then apply sampling. We use the name convention: `LogReg-<DST>-<sampling>(representation)`.

We tune all variations of DST and the baseline using 3-fold cross validation across documents per topic, where each fold is used as testing once. We tune the threshold $\theta \in [0.01, 0.02, ..., 1.0]$ and logistic regression parameter $\lambda \in \{0.001, 0.01, 0.1\}$ with grid search. Logistic regression is trained in a one-vs-rest scheme for handling multiple classes. We also tune all the class imbalance methods by making predictions with logistic regression on the new sample data after handling class imbalance: we tune the k-neighbours $k \in [2, 3, 5]$ and $out\_step \in [0.1, 0.5, 1.0]$ for Borderline Oversampling (BO), $k \in [2, 3, 5]$ for KmeansSmote, and $k \in [2, 3, 5]$ for Adasyn. All sampling approaches are applied only on the training data.

We use macro F1, averaged over all topics, to choose the best parameters. For inferring relevance labels, we report the best average performance over the 3 folds. For the other two experiments, we could not learn θ and λ for all the combinations of: document representations, sampling approaches, percentage of qrels and repetitions (2 700 combinations), as it was not computationally feasible. Thus we fix the θ and λ parameters to be the best average performance parameters used in the three folds.

Inferring Relevance Labels. We treat the problem of inferring labels as classification and train all models to predict if a document is non-relevant, relevant or highly relevant. Tab. 1 reports macro-F1 scores across the 5 TREC tracks. We see that, with few exceptions, TF-IDF is the best representation, followed by Glove and BERT, independently of the oversampling method. For BERT, most similarity values have a small effective range between 0.9-1.0, so the absolute difference in similarity between pairs is relatively small, which can obfuscate the signals necessary for the logistic regression to identify which pairs to give weights to.

We also see that DST is always beneficial with a mean gain of: 17.28% with min-max 0.9%-47.08% without any sampling, and 10.72% with min-max 0%-51.97% with oversampling. The threshold θ seems to effectively filter out low similarity documents which represent noise. Note that the maximum gain (%) is obtained when DST is used with BERT representation, as the θ parameter helps logistic regression to discriminate better between documents. Specifically, whenever the similarity pair is below the threshold θ, its similarity value is set to zero; this avoids the above problem of small effective range between similarities by increasing the range to 0-1.0.

Predicting system rankings and evaluation scores. To gauge the potential usefulness of the inferred labels, we perform two additional experiments: predicting the true ranking of systems (RoS) and predicting NDCG scores.

In the first task, the aim is to predict the RoS when labels of unjudged documents are replaced with inferred labels. We represent each TREC track by the RoS submitted to that track, and for each topic, we use the RoS produced using (a) the official and full TREC labels (qrels), and separately using (b) inferred labels and TREC official qrels with varying percentages. We compute the Rank-Biased Overlap (RBO) [23] coefficient with $\rho = 1.0$ between the two RoS for each topic and report its mean. The higher the correlation, the better the predicted RoS, thus the more accurate is the evaluation where labels of unjudged documents are replaced with inferred labels. We use RBO as it is top-heavy and can handle rankings with different elements (e.g. the top 10 systems per topic), while Kendall's τ and AP Correlation cannot [9, 24]. Tab. 1 shows RBO@10. [6] Overall, all models achieve high RBO values, suggesting that it is possible to infer the labels of unjudged documents, as per [4]. Document representation does not have a major impact on the prediction of the RoS. BERT is marginally the best representation, followed by Glove and TF-IDF. DST does not have a huge impact on the RoS prediction. It is beneficial for TF-IDF and Glove, but overall not for BERT. This may be because θ was not learned for all the combinations in Table 1.

For predicting evaluation scores, we analyze how accurately the inferred relevance labels of the documents approximate the original NDCG score. We report 1 minus the absolute difference between the average NDCG score of systems using the original and (partly) inferred document assessments. The higher the score, the better the NDCG score estimated when inferred labels are used.

Overall, all approaches perform well, with scores greater than 0.9 in most cases. This is aligned with the findings of [4], which show that the difference between the true MAP score and estimated MAP scores is greater than 0.88. All document representations are

[2]The source code is available online at URL-anonymized-for-review
[3]We use the 300-dimensional `glove6b` from https://nlp.stanford.edu/projects/glove/
[4]We use Uncased BERT-base from https://github.com/google-research/bert
[5]Online appendix: https://www.dropbox.com/s/12ghz00258xnham/CIKM2021_inferring_labels_appendix.pdf?dl=0
[6]Results on the whole RoS are included in the online appendix.

Table 1: Inferring labels: Macro F1-score on inferring graded relevance labels. Topics with documents assessed with only a single label (e.g. all documents assessed as non-relevant for a topic) are omitted. Predicting system rankings: Mean RBO@10 for each step of randomly removed documents from original qrels of the best 10 runs using NDCG per TREC track and the ranking produced using inferred document assessments. Each cell represents the mean across datasets and 100 repetitions. Predicting NDCG scores: Average 1 minus the absolute value of the difference between the original score using NDCG and the NDCG score obtained with the inferred labels for each step of randomly removed documents from original qrels. Each cell represents the mean across datasets and 100 repetitions.

Approach	Inferring labels					Predicting system rankings						Predicting NDCG scores					
	Web Track Collections					Web Track Collections 2010 - 2014						Web Track Collections 2010 - 2014					
						% of original qrels						% of original qrels					
	2010	2011	2012	2013	2014	85	70	55	40	25	10	85	70	55	40	25	10
Without Sampling																	
LogReg (TFIDF)	0.502	0.495	0.501	0.465	0.488	0.95	0.945	0.936	0.909	0.889	0.874	0.986	0.991	0.989	0.974	0.949	0.911
LogReg-DST (TFIDF)	0.524	0.515	0.526	**0.489**	**0.517**	0.97	**0.969**	0.937	0.903	0.896	0.856	0.986	0.989	0.987	0.972	0.947	0.91
LogReg (Glove)	0.499	0.486	0.531	0.419	0.449	0.963	0.948	0.938	0.892	0.88	0.849	0.988	0.989	0.98	0.962	0.935	0.895
LogReg-DST (Glove)	0.528	0.503	0.536	0.432	0.477	0.963	0.952	0.934	0.892	0.862	0.858	0.988	0.989	0.982	0.965	0.94	0.902
LogReg (BERT)	0.395	0.367	0.375	0.365	0.373	**0.983**	0.963	**0.959**	**0.950**	**0.921**	**0.924**	**0.989**	0.986	0.966	0.941	0.91	0.873
LogReg-DST (BERT)	**0.581**	**0.538**	**0.569**	0.483	0.513	0.963	0.952	0.904	0.893	0.865	0.852	0.986	**0.992**	**0.991**	**0.981**	**0.961**	**0.929**
Oversampling																	
LogReg-RO (TFIDF)	0.525	0.524	0.541	0.519	0.533	0.889	0.874	0.838	0.828	0.809	0.787	0.978	0.98	0.984	**0.989**	0.984	0.952
LogReg-DST-RO (TFIDF)	**0.542**	0.530	**0.560**	**0.527**	0.539	0.895	0.874	0.84	0.828	0.809	0.787	0.978	0.98	0.983	**0.989**	0.984	0.952
LogReg-BO (TFIDF)	0.514	0.517	0.530	0.521	0.535	0.952	0.939	0.92	0.906	0.88	0.864	0.967	0.955	0.943	0.929	0.918	0.911
LogReg-DST-BO (TFIDF)	0.538	**0.532**	0.548	0.522	**0.547**	0.952	0.934	0.92	0.902	0.88	0.837	0.967	0.955	0.943	0.93	0.918	0.911
LogReg-KmeansSmote (TFIDF)	0.502	0.494	0.512	0.470	0.508	0.954	**0.956**	**0.939**	0.906	0.899	**0.881**	**0.984**	**0.989**	**0.991**	0.986	0.969	0.936
LogReg-DST-KmeansSmote (TFIDF)	0.522	0.518	0.531	0.497	0.524	0.954	**0.956**	0.932	0.906	**0.902**	**0.881**	**0.984**	**0.989**	0.99	0.987	0.969	0.936
LogReg-adasyn (TFIDF)	0.506	0.509	0.526	0.502	0.532	**0.961**	0.935	0.922	0.911	0.89	0.867	0.978	0.979	0.982	0.988	**0.986**	**0.955**
LogReg-DST-adasyn (TFIDF)	0.532	0.527	0.544	0.515	0.540	**0.961**	0.937	0.918	**0.914**	0.88	0.867	0.978	0.979	0.982	0.988	**0.986**	**0.955**
LogReg-RO (Glove)	0.535	0.495	0.525	0.454	0.493	0.922	0.903	0.88	0.865	0.821	0.766	0.984	0.986	**0.987**	**0.984**	**0.966**	0.927
LogReg-DST-RO (Glove)	0.537	0.514	0.539	0.454	0.493	0.918	0.903	0.882	0.865	0.821	0.770	0.984	0.986	**0.987**	**0.984**	**0.966**	0.927
LogReg-BO (Glove)	0.520	0.495	0.521	0.447	0.484	0.957	0.957	0.939	0.922	0.879	0.852	0.98	0.979	0.974	0.968	0.962	**0.958**
LogReg-DST-BO (Glove)	**0.542**	**0.517**	**0.549**	**0.455**	**0.500**	0.957	0.957	0.939	0.917	0.875	0.852	0.98	0.979	0.974	0.968	0.962	**0.958**
LogReg-KmeansSmote (Glove)	0.498	0.493	0.512	0.419	0.453	**0.963**	**0.961**	0.946	0.933	0.91	0.898	**0.988**	**0.988**	0.983	0.971	0.949	0.913
LogReg-DST-KmeansSmote (Glove)	0.528	0.503	0.540	0.434	0.483	0.958	0.959	0.946	0.925	0.91	0.898	0.987	**0.988**	0.983	0.971	0.949	0.914
LogReg-adasyn (Glove)	0.507	0.494	0.507	0.439	0.470	0.957	0.953	**0.948**	**0.935**	**0.925**	**0.904**	0.984	0.986	0.986	0.983	0.964	0.925
LogReg-DST-adasyn (Glove)	0.526	0.507	0.538	0.448	0.486	0.957	0.953	**0.948**	**0.935**	**0.925**	**0.904**	0.984	0.986	0.986	0.983	0.964	0.925
LogReg-RO (BERT)	0.464	0.431	0.400	0.409	0.415	0.961	0.953	**0.947**	0.939	0.921	0.890	0.985	0.99	**0.991**	0.971	0.939	0.897
LogReg-DST-RO (BERT)	0.476	0.431	0.411	0.418	0.424	0.957	0.949	**0.947**	0.935	0.908	0.878	0.985	0.99	0.99	0.971	0.939	0.897
LogReg-BO (BERT)	0.409	0.391	0.381	0.390	0.401	**0.983**	0.963	0.932	0.932	0.921	0.919	0.98	0.983	0.989	**0.992**	**0.971**	**0.924**
LogReg-DST-BO (BERT)	**0.583**	**0.547**	**0.579**	**0.492**	0.523	**0.983**	0.943	0.932	0.932	0.921	0.914	0.98	0.983	0.988	**0.992**	**0.971**	**0.924**
LogReg-KmeansSmote (BERT)	0.406	0.386	0.382	0.374	0.389	**0.983**	**0.978**	0.934	0.932	0.92	0.919	**0.989**	0.989	0.973	0.95	0.919	0.879
LogReg-DST-KmeansSmote (BERT)	0.578	0.535	0.565	0.485	0.389	**0.983**	**0.978**	0.934	0.932	0.92	0.908	**0.989**	0.989	0.973	0.95	0.919	0.879
LogReg-adasyn (BERT)	0.441	0.419	0.401	0.400	0.405	0.959	0.957	0.945	**0.944**	**0.948**	**0.932**	**0.989**	**0.991**	0.974	0.949	0.916	0.876
LogReg-DST-adasyn (BERT)	0.574	0.544	0.570	**0.492**	**0.524**	0.959	0.953	0.945	0.943	0.943	**0.932**	**0.989**	**0.991**	0.974	0.949	0.915	0.876

also on par. When combined with oversampling, BERT and Glove seem marginally better than TF-IDF. DST does not seem to have an impact on the prediction of NDCG scores, neither when combined with sampling approaches, nor when considered in isolation. As for the previous case, θ was not learned for each combination of models and repetition of this experiment, this might affect DST performance. We see that all oversampling methods improve over the baseline for all representations when using less than 70% of the original qrels. BO, Kmeans, and ADASYN achieve the best scores: BO and kmeans have comparable results when used with BERT, ADASYN seems to work well in the label-scarce scenario (lower than 55% of original qrels).

5 CONCLUSION

We study the problem of inferring relevance labels for unjudged documents in IR datasets. Currently, this is done without accounting for the imbalance between relevant (very few) and non-relevant (many) documents, which means that the discriminative signal of relevant documents (minority) is weakened by the much stronger noise signal of non-relevant documents (majority). We present a simple method, called DST, for reducing the noisy impact of highly dissimilar documents when inferring relevance assessments from document similarities. This is done by defining a decision boundary θ to mitigate the noisy signal contributed by the high number of dissimilar (non-relevant) documents. Experiments on 5 TREC tracks (248 topics, 256 runs), where we apply DST either alone, or combined with oversampling, show that DST is effective (both without and with oversampling). Albeit simple, our method yields a mean gain of +17.28% without oversampling and +10.72% with oversampling, which means that there is room for improvement. This paves the way for future work in refining alternatives to the regression model, and establishing more topologically sensitive boundary conditions for the decision boundary rather than the current single threshold.

REFERENCES

[1] Javed A. Aslam and Robert Savell. 2003. On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 361–362. https://doi.org/10.1145/860435.860501

[2] Javed A. Aslam and Emine Yilmaz. 2007. Inferring Document Relevance from Incomplete Information. In *Proceedings of CIKM 2007*. Association for Computing Machinery, New York, NY, USA, 633–642.

[3] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the limits of pooling for large collections. *Information Retrieval* 10, 6 (2007), 491–508.

[4] Ben Carterette and James Allan. 2007. Semiautomatic Evaluation of Retrieval Systems Using Document Similarities. In *Proceedings of CIKM 2007*. ACM, 873–876.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*. ACL, 4171–4186.

[7] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 1322–1328.

[8] N. Jardine and C.J. van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7, 5 (1971), 217 – 240.

[9] M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (November 1945), 239–251.

[10] Felix Last, Georgios Douzas, and Fernando Bação. 2017. Oversampling for Imbalanced Learning Based on K-Means and SMOTE. (11 2017).

[11] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. 2017. Clustering-based undersampling in class-imbalanced data. *Information Sciences* 409-410 (2017), 17 – 26. https://doi.org/10.1016/j.ins.2017.05.008

[12] Stuart P. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory* 28 (1982), 129–137.

[13] Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. 2011. Borderline oversampling for imbalanced data classification. *IJKESDP* 3, 1 (2011), 4–21.

[14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*. 1532–1543.

[15] Kevin Roitero, Andrea Brunello, Giuseppe Serra, and Stefano Mizzaro. 2020. Effectiveness Evaluation Without Human Relevance Judgments: A Systematic Analysis of Existing Methods and of their Combinations. *Information Processing & Management* 57, 2 (2020), 102–149.

[16] Tetsuya Sakai. 2009. On the Robustness of Information Retrieval Metrics to Biased Relevance Assessments. *Information and Media Technologies* 4, 2 (2009), 547–557.

[17] Tetsuya SAKAI, Noriko Kando, Hideki Shima, Chuan-Jie Lin, Ruihua Song, Miho Sugimoto, and Teruko Mitamura. 2009. Ranking the NTCIR ACLIA IR4QA Systems without Relevance Assessments. *Information and Media Technologies* 4, 4 (2009), 1028–1033. https://doi.org/10.11185/imt.4.1028

[18] Michael R. Smith, Tony Martinez, and Christophe Giraud-Carrier. 2014. An Instance Level Analysis of Data Complexity. *Mach. Learn.* 95, 2 (May 2014), 225–256. https://doi.org/10.1007/s10994-013-5422-z

[19] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 66–73. https://doi.org/10.1145/383952.383961

[20] Anselm Spoerri. 2007. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Processing & Management* 43 (07 2007), 1059–1070. https://doi.org/10.1016/j.ipm.2006.09.009

[21] Ivan Tomek. 1976. An Experiment with the Edited Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6, 6 (1976), 448–452.

[22] C. J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

[23] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (Nov. 2010), 38 pages. https://doi.org/10.1145/1852102.1852106

[24] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. 2008. A New Rank Correlation Coefficient for Information Retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 587–594. https://doi.org/10.1145/1390334.1390435

Bibliography

[1] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012.

[2] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas, November 2016. Association for Computational Linguistics.

[3] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset, 2017.

[4] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, page 5–14, New York, NY, USA, 2009. Association for Computing Machinery.

[5] Ameer Albahem, Damiano Spina, Falk Scholer, and Lawrence Cavedon. Meta-evaluation of dynamic search: How do metrics capture topical relevance, diversity and user effort? In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, pages 607–620, Cham, 2019. Springer International Publishing.

[6] Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. Time-aware evidence ranking for fact-checking. *arXiv preprint arXiv:2009.06402*, 2020.

[7] E. Amigó and S. Mizzaro. On the Nature of Information Access Evaluation Metrics: a Unifying Framework. *Information Retrieval Journal*, 23(3):318–386, 2020.

[8] Enrique Amigó, Fernando Giner, Stefano Mizzaro, and Damiano Spina. A formal account of effectiveness evaluation and ranking fusion. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '18, page 123–130, New York, NY, USA, 2018. Association for Computing Machinery.

[9] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, SIGIR '18, page 625–634, New York, NY, USA, 2018. Association for Computing Machinery.

[10] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 27–34, 2005.

[11] Javed A. Aslam, Virgiliu Pavlu, and Robert Savell. A unified model for metasearch, pooling, and system evaluation. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, page 484–491, New York, NY, USA, 2003. Association for Computing Machinery.

[12] Javed A. Aslam and Emine Yilmaz. Inferring document relevance from incomplete information. In *Proceedings of CIKM 2007*, page 633–642, New York, NY, USA, 2007. Association for Computing Machinery.

[13] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3256–3274. Association for Computational Linguistics, 2020.

[14] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 7352–7364. Association for Computational Linguistics, 2020.

[15] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China, 2019. Association for Computational Linguistics.

[16] Ricardo Baeza-Yates. Applications of web query mining. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval*, pages 7–22, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[17] Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics, 2003.

[18] Ramy Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating Stance Detection and Fact Checking in a Unified Corpus. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 21–27. Association for Computational Linguistics, 2018.

[19] Carol L. Barry and Linda Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing Management*, 34(2):219–236, 1998.

[20] Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018.

[21] Pia Borlund. The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003.

[22] S. Borzsony, D. Kossmann, and K. Stocker. The skyline operator. In *Proceedings 17th International Conference on Data Engineering*, pages 421–430, 2001.

[23] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen M. Voorhees. Bias and the limits of pooling for large collections. *Inf. Retr.*, 10(6):491–508, 2007.

[24] Chris Buckley and Ellen Voorhees. Retrieval system evaluation. *TREC: Experiment and Evaluation in Information Retrieval*, pages 53–75, 01 2005.

[25] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 33–40, New York, NY, USA, 2000. ACM.

[26] Vannevar Bush. As We May Think. *Atlantic Monthly*, 176(1):641–649, March 1945.

[27] Luca Busin and Stefano Mizzaro. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, page 22–29, New York, NY, USA, 2013. Association for Computing Machinery.

[28] Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[29] Ben Carterette and James Allan. Semiautomatic evaluation of retrieval systems using document similarities. In Mário J. Silva, Alberto H. F. Laender, Ricardo A. Baeza-Yates, Deborah L. McGuinness, Bjørn Olstad, Øystein Haug Olsen, and André O. Falcão, editors, *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 873–876. ACM, 2007.

[30] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[31] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 621–630, New York, NY, USA, 2009. Association for Computing Machinery.

[32] Lucas Chaves Lima, Casper Hansen, Christian Hansen, Dongsheng Wang, Maria Maistro, Birger Larsen, Jakob Grue Simonsen, and Christina Lioma. Denmark's Participation in the Search Engine TREC COVID-19 Challenge: Lessons Learned about Searching for Precise Biomedical Scientific Information on COVID-19. *arXiv e-prints*, page arXiv:2011.12684, November 2020.

[33] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 15–24, New York, NY, USA, 2017. ACM.

[34] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[35] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 659–666. ACM Press, New York, USA, 2008.

[36] Charles Clarke, Maria Maistro, Mark Smucker, and Guido Zuccon. Overview of the TREC 2020 Health Misinformation Track (to appear). In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Nine Text REtrieval Conference, TREC 2020, Gaithersburg, Maryland,*

USA, November 16-19, 2020, volume - of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020.

[37] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2009 web track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, volume 500-278 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2009.

[38] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. Overview of the TREC 2010 web track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*, volume 500-294 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2010.

[39] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. Overview of the TREC 2011 web track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011*, volume 500-296 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2011.

[40] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. Overview of the TREC 2012 web track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, volume 500-298 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2012.

[41] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 659–666, New York, NY, USA, 2008. Association for Computing Machinery.

[42] C.W. Cleverdon, J. Mills, Aslib. Cranfield Research Project, and M. Keen.

Factors Determining the Performance of Indexing Systems. Number vb. 1 in Factors Determining the Performance of Indexing Systems. College of Aeronautics, 1966.

[43] Cyril W. Cleverdon, J. Mills, and Michael Keen. *Factors determining the performance of indexing systems*. College of Aeronautics, 1966.

[44] P. Clough and M. Sanderson. Evaluating the performance of information retrieval systems using test collections. *Information Research*, 18(2), June 2013. © Year The Author(s). This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[45] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. TREC 2013 web track overview. In Ellen M. Voorhees, editor, *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, volume 500-302 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2013.

[46] Kevyn Collins-Thompson, Craig Macdonald, Paul N. Bennett, Fernando Diaz, and Ellen M. Voorhees. TREC 2014 web track overview. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, volume 500-308 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2014.

[47] William S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971.

[48] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 282–289, New York, NY, USA, 1998. Association for Computing Machinery.

[49] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language under-

standing. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, pages 4171–4186. Association for Computational Linguistics, 2019.

[50] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[51] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. Checkthat! at clef 2019: Automatic identification and verification of claims. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, pages 309–315, Cham, 2019. Springer International Publishing.

[52] Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. Towards linguistically generalizable NLP systems: A workshop and shared task. *CoRR*, abs/1711.01505, 2017.

[53] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. A multiword expression dataset: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions (MWE-NAACL 2015)*. Association for Computational Linguistics, 2015.

[54] M. Ferrante, N. Ferro, and E. Losiouk. How do Interval Scales Help us with Better Understanding IR Evaluation Measures? *Information Retrieval Journal*, 2019.

[55] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. Towards meaningful statements in IR evaluation. mapping evaluation measures to interval scales. *CoRR*, abs/2101.02668, 2021.

[56] Marco Ferrante, Nicola Ferro, and Maria Maistro. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proceedings of the 2015 International Conference on The Theory of Information*

Retrieval, ICTIR '15, page 21–30, New York, NY, USA, 2015. Association for Computing Machinery.

[57] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1161–1166. Association for Computational Linguistics, 2019.

[58] H. Gilbert and K. Jones. Statistical bases of relevance assessment for the ideal information retrieval test collection. 1979.

[59] Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Inf. Fusion*, 66:184–197, 2021.

[60] Yoav Goldberg and Graeme Hirst. *Neural Network Methods in Natural Language Processing*. Morgan amp; Claypool Publishers, 2017.

[61] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[62] Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. Contextually propagated term weights for document representation. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 897–900. ACM, 2019.

[63] Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. Neural check-worthiness ranking with weak supervi-

sion: Finding sentences for fact-checking. In Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 994–1000. ACM, 2019.

[64] Casper Hansen, Christian Hansen, and Lucas Chaves Lima. Automatic fake news detection: Are models learning to reason? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2021. Association for Computational Linguistics.

[65] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. Unsupervised neural generative semantic hashing. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 735–744. ACM, 2019.

[66] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. Content-aware neural hashing for cold-start recommendation. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 971–980. ACM, 2020.

[67] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. Unsupervised semantic hashing with pairwise reconstruction. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2009–2012. ACM, 2020.

[68] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma. Fact check-worthiness detection with contrastive ranking. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis,

Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 124–130. Springer, 2020.

[69] Christian Hansen, Casper Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. Unsupervised multi-index semantic hashing. *Proceedings of the 2021 World Wide Web Conference*, abs/2103.14460, 2021.

[70] Christian Hansen, Casper Hansen, Jakob Grue Simonsen, Birger Larsen, Stephen Alstrup, and Christina Lioma. Factuality checking in news headlines with eye tracking. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2013–2016. ACM, 2020.

[71] Donna K Harman, EM Voorhees, et al. *The Text REtrieval Conference (TREC)*. US Department of Commerce, National Bureau of Standards, 1993.

[72] Kazuma Hashimoto and Yoshimasa Tsuruoka. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 205–215, Berlin, Germany, August 2016. Association for Computational Linguistics.

[73] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812, 2017.

[74] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR

'00, page 17–24, New York, NY, USA, 2000. Association for Computing Machinery.

[75] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[76] Christopher Horn, Alisa Zhila, Alexander Gelbukh, Roman Kern, and Elisabeth Lex. Using factual density to measure informativeness of web documents. 01 2013.

[77] Lee Howell et al. Digital wildfires in a hyperconnected world. *WEF report*, 45(3):15–94, 2013.

[78] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, S. M. Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. Covid-19?related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621 – 1629, 07 Oct. 2020.

[79] Jim Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing Management*, 36:207–227, 03 2001.

[80] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[81] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 41–48, New York, NY, USA, 2000. Association for Computing Machinery.

[82] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[83] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods*

in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 2021–2031. Association for Computational Linguistics, 2017.

[84] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[85] N. Kando, editor. *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*. National Center for Science Information Systems, Published Online, 1999.

[86] Graham Katz and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Association for Computational Linguistics, 2006.

[87] Douwe Kiela and Stephen Clark. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1427–1432. Association for Computational Linguistics, 2013.

[88] Eva L. Kiewitt. *Evaluating information retrieval systems, the PROBE program / Eva L. Kiewitt ; foreword by Bernard M. Fry*. Greenwood Press Westport, Conn, 1979.

[89] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[90] Klas Leino, Matt Fredrikson, Emily Black, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. In *International Conference on Learning Representations*, 2019.

[91] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do supervised distributional methods really learn lexical inference relations? In *Pro-*

ceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 970–976, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[92] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[93] Lucas Chaves Lima, Casper Hansen, Christian Hansen, Maria Maistro, Jakob Grue Simonsen, and Christina Lioma. Test collection incompleteness and unjudged documents (under revision), 2020.

[94] Lucas Chaves Lima, Casper Hansen, Christian Hansen, Dongsheng Wang, Maria Maistro, Birger Lansen, Jakob Grue Simonsen, and Christina Lioma. Denmark's participation in the search engine trec covid-19 challenge: Lessons learned about searching for precise biomedical scientific information on covid-19 (to appear). In *Proceedings of the Twenty-Nine Text REtrieval Conference, TREC 2020, Gaithersburg, Maryland, USA, November 16-19, 2020*, volume - of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020.

[95] Lucas Chaves Lima, Dustin Wright, Isabelle Augestein, and Maria Maistro. University of copenhagen participation in trec health misinformation track 2020 (to appear). In *Proceedings of the Twenty-Nine Text REtrieval Conference, TREC 2020, Gaithersburg, Maryland, USA, November 16-19, 2020*, volume - of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020.

[96] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[97] C. Lioma, M. Maistro, M. D. Smucker, , and D. Zuccon. Overview of the TREC 2019 Decision Track.

[98] Christina Lioma. Dependencies: Formalising semantic catenae for information retrieval. *CoRR*, abs/1709.03742, 2017.

[99] Christina Lioma and Niels Dalum Hansen. A study of metrics of distance and correlation between ranked lists for compositionality detection. *Cogn. Syst. Res.*, 44(C):40–49, August 2017.

[100] Christina Lioma, Birger Larsen, and Peter Ingwersen. Preliminary experiments using subjective logic for the polyrepresentation of information needs. In Jaap Kamps, Wessel Kraaij, and Norbert Fuhr, editors, *Information Interaction in Context: 2012, IIix'12, Nijmegen, The Netherlands, August 21-24, 2012*, pages 174–183. ACM, 2012.

[101] Christina Lioma, Birger Larsen, Wei Lu, and Yong Huang. A study of factuality, objectivity and relevance: three desiderata in large-scale information retrieval? In Ashiq Anjum and Xinghui Zhao, editors, *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, BDCAT 2016, Shanghai, China, December 6-9, 2016*, pages 107–117. ACM, 2016.

[102] Christina Lioma, Birger Larsen, Hinrich Schütze, and Peter Ingwersen. A subjective logic formalisation of the principle of polyrepresentation for information needs. In Nicholas J. Belkin and Diane Kelly, editors, *Information Interaction in Context Symposium, IIiX 2010, New Brunswick, NJ, USA, August 18-21, 2010*, pages 125–134. ACM, 2010.

[103] Christina Lioma and Iadh Ounis. Light syntactically-based index pruning for information retrieval. In Giambattista Amati, Claudio Carpineto, and Giovanni Romano, editors, *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, volume 4425 of *Lecture Notes in Computer Science*, pages 88–100. Springer, 2007.

[104] Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. Evaluation measures for relevance and credibility in ranked lists. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 91–98, New York, NY, USA, 2017. ACM.

[105] Christina Lioma, Jakob Grue Simonsen, Birger Larsen, and Niels Dalum Hansen. Non-compositional term dependence for information retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–604. ACM, 2015.

[106] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[107] David E. Losada, Javier Parapar, and Alvaro Barreiro. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing Management*, 53(5):1005–1025, 2017.

[108] Xiaolu Lu, Alistair Moffat, and J. Shane Culpepper. The effect of pooling and evaluation depth on ir metrics. *Inf. Retr.*, 19(4):416–445, August 2016.

[109] Joel M. Mackenzie, Zhuyun Dai, Luke Gallagher, and Jamie Callan. Efficiency implications of term weighting for passage retrieval. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1821–1824. ACM, 2020.

[110] E. Maddalena and S. Mizzaro. Axiometrics: Axioms of information retrieval effectiveness metrics. In Cranefield S. Trotman, A. and J. Yang, editors, *Australasian Web Conference (AWC 2014)*, volume 155 of *CRPIT*, pages 39–48, Auckland, New Zealand, 2014. ACS.

[111] Maria Maistro, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. Principled multi-aspect evaluation measures of rankings (under revision), 2020.

[112] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008.

[113] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024, 2019.

[114] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the*

ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09, page 309–312, USA, 2009. Association for Computational Linguistics.

[115] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[116] Alistair Moffat. Seven numeric properties of effectiveness metrics. In Rafael E. Banchs, Fabrizio Silvestri, Tie-Yan Liu, Min Zhang, Sheng Gao, and Jun Lang, editors, *Information Retrieval Technology*, pages 1–12, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[117] Alistair Moffat, Falk Scholer, and Paul Thomas. Models and metrics: Ir evaluation as a user process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, ADCS '12, pages 47–54, New York, NY, USA, 2012. ACM.

[118] Alistair Moffat, William Webber, and Justin Zobel. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 375–382, New York, NY, USA, 2007. Association for Computing Machinery.

[119] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December 2008.

[120] Nafise Sadat Moosavi and Michael Strube. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[121] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar, October 2014. Association for Computational Linguistics.

[122] Joao Palotti, Guido Zuccon, and Allan Hanbury. Mm: A new framework for multidimensional evaluation of search engines. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1699–1702, New York, NY, USA, 2018. Association for Computing Machinery.

[123] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[124] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[125] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[126] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics, 2018.

[127] C. Peters, F. C. Gey, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke, editors. *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers*. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany, 2006.

[128] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[129] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, 2018.

[130] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[131] Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. Defactonlp: Fact verification using entity recognition, TFIDF vector comparison and decomposable attention. *CoRR*, abs/1809.00509, 2018.

[132] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[133] Kevin Roitero, Andrea Brunello, Giuseppe Serra, and Stefano Mizzaro. Effectiveness Evaluation Without Human Relevance Judgments: A Systematic Analysis of Existing Methods and of their Combinations. *Information Processing & Management*, 57(2):102–149, 2020.

[134] T. Sakai. Evaluating information retrieval metrics based on bootstrap hypothesis tests. *IPSJ Digital Courier*, 3:625–642, 2007.

[135] Tetsuya Sakai. On the robustness of information retrieval metrics to biased relevance assessments. *Information and Media Technologies*, 4(2):547–557, 2009.

[136] Tetsuya Sakai. Evaluation with informational and navigational intents. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 499–508, New York, NY, USA, 2012. Association for Computing Machinery.

[137] Bahar Salehi, Paul Cook, and Timothy Baldwin. Detecting non-compositional mwe components using wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1792–1797, 2014.

[138] Bahar Salehi, Paul Cook, and Timothy Baldwin. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[139] Bahar Salehi, Paul Cook, and Timothy Baldwin. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, 2015.

[140] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375, 2010.

[141] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 555–562, New York, NY, USA, 2010. ACM.

[142] Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[143] Julia Schwarz and Meredith Ringel Morris. Augmenting web pages and search results to support credibility assessment. In *ACM Conference on Computer-Human Interaction*, May 2011.

[144] Rico Sennrich and Barry Haddow. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 83–91. The Association for Computer Linguistics, 2016.

[145] Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. Additive compositionality of word vectors. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 387–396, Hong Kong, China, November 2019. Association for Computational Linguistics.

[146] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, 2020.

[147] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[148] Louise T. Su. Evaluation measures for interactive information retrieval. *Information Processing and Management*, 28(4):503 – 516, 1992. Special Issue: Evaluation Issues in Information Retrieval.

[149] J. Tague-Sutcliffe and J. Blustein. A statistical analysis of the trec-3 data. In *TREC*, 1994.

[150] Jean Tague-Sutcliffe and James Blustein. A statistical analysis of the TREC-3 data. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume Special Publication 500-225, page 385, Maryland, USA, 1994. National Institute of Standards and Technology (NIST).

[151] Z. Tang and G. H. Yang. Investigating per topic upper bound for session search evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, page 185–192, New York, NY, USA, 2017. Association for Computing Machinery.

[152] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[153] Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 11–18, New York, NY, USA, 2006. ACM.

[154] Sagar Uprety, Yi Su, Dawei Song, and Jingfei Li. Modeling multidimensional user relevance in IR using vector spaces. *CoRR*, abs/1805.02184, 2018.

[155] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[156] Sriram Venkatapathy and Aravind K Joshi. Measuring the relative compositionality of verb-noun (vn) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 899–906. Association for Computational Linguistics, 2005.

[157] Manisha Verma, Emine Yilmaz, Rishabh Mehrotra, Evangelos Kanoulas, Ben Carterette, Nick Craswell, and Peter Bailey. Overview of the TREC tasks track 2016. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, volume 500-321 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2016.

[158] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5797–5808. Association for Computational Linguistics, 2019.

[159] Ellen M. Voorhees. Overview of TREC 2001. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001*, volume 500-250 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2001.

[160] Ellen M. Voorhees. Overview of TREC 2002. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Eleventh Text REtrieval Con-*

ference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002, volume 500-251 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2002.

[161] Ellen M. Voorhees. Overview of TREC 2003. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003*, volume 500-255 of *NIST Special Publication*, pages 1–13. National Institute of Standards and Technology (NIST), 2003.

[162] Ellen M. Voorhees. Overview of TREC 2004. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2004.

[163] Ellen M. Voorhees. Overview of TREC 2007. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, volume 500-274 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2007.

[164] Ellen M. Voorhees. On test collections for adaptive information retrieval. *Information Processing Management*, 44(6):1879–1885, 2008. Adaptive Information Retrieval.

[165] Ellen M. Voorhees and Donna Harman. Overview of the eighth text retrieval conference (TREC-8). In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 1999.

[166] Ellen M. Voorhees and Donna Harman. Overview of the ninth text retrieval conference (TREC-9). In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*, volume 500-249 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2000.

[167] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[168] Dongsheng Wang, Casper Hansen, Lucas Chaves Lima, Christian Hansen, Maria Maistro, Jakob Grue Simonsen, and Christina Lioma. Multi-head self-attention with role-guided masks (under review), 2020.

[169] Dongsheng Wang, Qiuchi Li, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. Contextual compositionality detection with external knowledge bases and word embeddings. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 317–323, New York, NY, USA, 2019. Association for Computing Machinery.

[170] Dongsheng Wang, Qiuchi Li, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. Contextual compositionality detection with external knowledge bases and word embeddings. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 317–323, New York, NY, USA, 2019. Association for Computing Machinery.

[171] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426. Association for Computational Linguistics, 2017.

[172] W. Webber, A. Moffat, J. Zobel, and T. Sakai. Precision-at-ten considered redundant. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 695–696, New York, NY, USA, 2008. ACM.

[173] William Webber. *Measurement in information retrieval evaluation*. PhD thesis, University of Melbourne, Australia, 2010.

[174] Yunjie (Calvin) Xu and Zhiwei Chen. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, 57(7):961–973, 2006.

[175] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. Commonly uncommon: Semantic sparsity in situation recognition, 2016.

[176] Majid Yazdani, Meghdad Farahmand, and James Henderson. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, 2015.

[177] Emine Yilmaz and Stephen Robertson. On the choice of effectiveness measures for learning to rank. *Information Retrieval Journal*, 13(3):271 – 290, 2010.

[178] Emine Yilmaz, Manisha Verma, Rishabh Mehrotra, Evangelos Kanoulas, Ben Carterette, and Nick Craswell. Overview of the TREC 2015 tasks track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.

[179] Haotian Zhang. *Increasing the Efficiency of High-Recall Information Retrieval*. PhD thesis, University of Waterloo, Ontario, Canada, 2019.

[180] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[181] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[182] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 307–314, New York, NY, USA, 1998. Association for Computing Machinery.

[183] Guido Zuccon. Understandability biased evaluation for information retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, pages 280–292, Cham, 2016. Springer International Publishing.