Entity Detection for Check-worthiness Prediction: Glasgow Terrier at CLEF CheckThat! 2019

Ting Su[1], Craig Macdonald[2], and Iadh Ounis[2]

University of Glasgow, Glasgow, UK
[1] t.su.2@research.gla.ac.uk
[2] firstname.lastname@glasgow.ac.uk

Abstract. Since information can be created and shared online by anyone, a lot of time and effort are required to manually fact-check all the information encountered by users everyday. Hence, an automatic fact-checking process is needed to effectively fact-check the vast information available online. However, gathering information related to every single claim can also be redundant, as not all sentences or articles are check-worthy. In this paper, we propose an effective approach for retrieving check-worthy sentences within American political debates, which relates to the first task of the CLEF CheckThat! 2019 Lab. To rank sentences based on their check-worthiness, we propose to represent each sentence using their mentioned entities using a TF-IDF representation. We use a SVM classifier to predict the check-worthiness of each sentence. Our approach ranked 4th out of 12 submissions. Our experiments show that the pronouns and coreference resolution pre-processing procedure we use as part of our approach does improve the effectiveness of sentence check-worthiness prediction. Furthermore, our results show that entity analysis features provide valuable evidence for this task.

Keywords: Fact checking · Entity relationships · Check-worthiness

1 Introduction

Nowadays, information is easily accessible online, from articles by reliable news agencies, to reports from independent reporters, to extreme views published by unknown individuals. Such amount of information may create difficulties for information consumers as they try to distinguish fake news from genuine news. Indeed, users may not be necessarily aware that the information they encounter is false, and may not have the time and effort to fact-check all the claims and information they come across online. Moreover, social media outlets are becoming increasingly important in everyday life, where users can obtain the latest news

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

CEUR-WS.org/Vol-2380/paper_109.pdf

Table 1. Examples of sentences to check. **Bold** denotes entities.

Hillary Clinton:	I think my husband did a pretty good job in the 1990s.	
Hillary Clinton:	I think a lot about what worked and how we can make it work again...	
Donald Trump:	Well, he approved **NAFTA**..	✓
Hillary Clinton:	Take **clean energy**.	
Hillary Clinton:	**Donald** thinks that **climate change** is a hoax perpetrated by **china**	✓
Hillary Clinton:	I think it's real.	
Donald Trump:	I did not.	

and updates, share links to news and information they want to spread, and post comments with their own opinions. With the amount of information that is created daily, it is not feasible for journalists and users to manually fact-check every news article, sentence or tweet online. Therefore, an automatic fact-checking system that extracts the most check-worthy claims from articles and debates could allow journalists to focus on manually checking suspicious but worthy claims, thereby reducing the workload required for the task.

The task of predicting the check-worthiness of each sentence in the text is the objective of Task 1 of the CLEF CheckThat! 2019 Lab [1]. In particular, participants are asked to retrieve the most check-worthy sentences from transcripts obtained from American political debates. The task is defined as follows. Given a debate (D) that contains a set of ordered sentences ($D = (s_1, s_2, ..., s_3)$), where each sentence has a line number, a speaker's name, and the content of the sentence ($s_n =< l_n, p_n, c_n >$)), a system should return a list of sentences, ordered based on their estimated check-worthiness. For example, Table 1 presents two examples of excerpts from such debates, where the sentences labelled with ✓ are considered to be check-worthy.

The focus of this paper is to effectively address Task 1 of the CLEF CheckThat! 2019 Lab. To do so, we build upon recent developments to improve a chatbot's understanding in a conversation (a debate is form of conversation [3]), namely techniques for coreference resolution, in order to process the pronouns present within the text. Moreover, we observe that several entities tend to be present in sentences that are worth checking. For example, the bold text in Table 1 refers to entities. Therefore, we hypothesise that an entity resolution and analysis using knowledge graphs (KG) can help distinguish between sentences that are worth checking and sentences that are not. The contributions of this paper are two-fold: we develop a useful automatic pre-processing procedure to process the text before analysis; Secondly, we show that entity resolution and analysis can indeed enhance the effectiveness of our approach at identifying check-worthy sentences.

The rest of the paper is organised as follows. We briefly introduce related work in Section 2. Section 3 describes our proposed approach. We provide the experimental setup in Section 4, followed by the results and analysis in Section 5. Finally, we draw the main conclusions from this paper in Section 6.

2 Related Work

Previous studies that focused on the task of predicting the check-worthiness of a sentence are limited. ClaimBuster [6] used an SVM classifier with TF-IDF, part-of-speech (POS) tags, and named entity recognition as features, to classify a sentence into *factual*, *unimportant-factual*, and *check-worthy factual*. Gencheva et al., [4] improved the work of ClaimBuster by using additional sentiment, tense, and paragraph structure features, to predict if a sentence should be fact-checked. This work was further improved, and resulted in ClaimRank [8], which can provide journalists with check-worthy sentences for manual checking. Moreover, Patwari et al., [10] used an SVM classifier to classify if a sentence is check-worthy or not. In particular, they analysed the topic a sentence is talking about using an LDA topic modelling approach. They also used POS with TF-IDF representation features, and achieved a 0.214 F1 score.

In last year's CLEF CheckThat! Lab, aside from the above mentioned methods, team Prise de Fer [14] manually normalised names and pronouns that appeared in the debate as a pre-processing procedure. They also used clauses and phrases as well as rule-based heuristics on the length of the sentence within a multilayer perceptron. Team Copenhagen [5] used word2vec embedding and a recurrent neural network model, and achieved 0.182 in mean average precision in a check-worthy sentence retrieval task. In addition, syntactic dependencies were used by both teams [5, 14].

However, the above mentioned approaches did not pay much attention to *automatic* pre-processing, in order to unify the pronouns and references. Moreover, although these approaches used named entities as features, none of these approaches used external resources to analyse the entities mentioned in the text. Our work focuses on these two parts of analysing sentences, to predict their check-worthiness.

3 Our Entity Detection Approach

The aim of the check-worthiness task is to rank the sentences, such that those sentences estimated most likely to be check-worthy are ranked first. In addressing this task, we use a classifier based on several groups of features to estimate the check-worthiness of each sentence. Sentences are then ranked based on the classifier's confidence about the check-worthiness of each sentence. Our classification approach makes use of a pre-processing of the text that addresses pronouns and coreference resolution (described in detail in Section 3.1 below), as well as several groups of features, including some that consider the presence of entities within the sentences (Section 3.2).

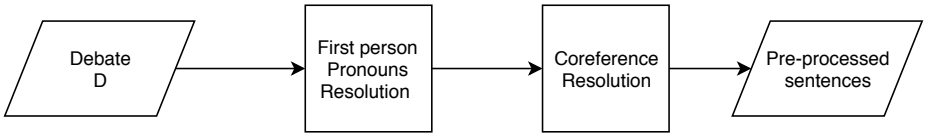

Fig. 1. The pre-processing procedure. A parallelogram represents input and/or output; a rectangle represents a process; an arrow represents the relationship flow between two components.

3.1 Pre-processing Procedure

American political debates usually consist of two or more participants, and one or more moderators, where each debate has different participants. In this case, it is not explicitly apparent to the system which participants are referenced by which pronouns. Similarly, implicit pronouns can also be used to identify a specific person or a particular thing previously mentioned or known, leading to a possible confusion. To combat the above mentioned challenges in implicit references, we propose a two-step procedure to resolve the implicit references found in the debates, namely, first-person pronouns resolution, and coreference resolution, as illustrated in Figure 1. Detailed examples can be found in Table 5.

1. First-person pronouns resolution: In this step, we simply change all the first-person pronouns in each sentence s_n into the current speaker's name p_n.

2. Coreference resolution: Coreference resolution is the task of finding the entity expression that a pronoun refers to within a piece of text. In our proposed procedure, we use coreference resolution to replace implicit mentions to one of the previously stated real-world entities. Specifically, we use Lee et al. [9]'s implementation of a higher order coreference resolution method, applies on pairs of sentences. Therefore, the span of possible references for a pronoun is from either the current sentence, or the antecedent of the sentence, regardless of any change in speaker.

3.2 Check-worthiness Estimation

After the pre-processing procedure, we obtain an ordered (based on the order of the debate) list of sentences for each debate, where most pronouns are replaced with the person's name and/or entities. Note that the coreference resolution method cannot achieve a perfect accuracy, as only a subset of the actual pronouns are resolved.

Next, to obtain a ranking of sentences, we extract features from each sentence, and use these features as input to a classifier that is trained to estimate the check-worthiness of each sentence. Figure 2 shows the overall architecture of our proposed approach. Below, we describe two sources of features that we use to assist the check-worthiness estimation, namely TF-IDF sentence representation and entities analysis.

1. TF-IDF sentence representation: We calculate the TF-IDF score of each term in each sentence, where the IDF values are calculated over all of the

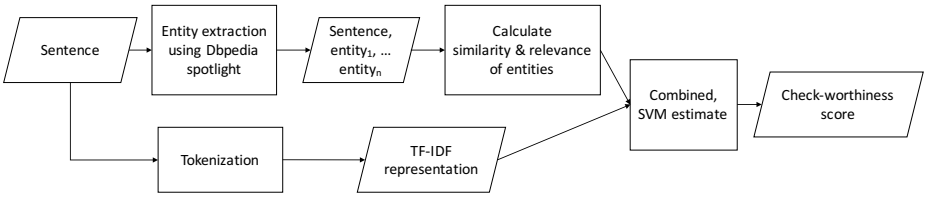

Fig. 2. The check-worthiness estimation approach. A parallelogram represents input and/or output; a rectangle represents a process; an arrow represents the relationship flow between two components.

training set. In particular, we use Sklearn's TfidfVectorizer[1] to extract features for each sentence. We do not discard any terms from the dictionary.

2. Entity analysis: Our second group of features concerns the entities that appear in each sentence, obtained through entity linking occurrences in each sentence to a knowledge graph (KG), namely Wikipedia. In particular, Wikipedia is a large-scale online encyclopedia, where users can create articles related to specific entities, and can edit existing articles. The crowd-sourcing nature of Wikipedia allows the entities' information to be updated quickly, which means that the information is kept up-to-date. Wikipedia also contains structure relationships where one or more entities are linked together through hyperlinks, such as Polysemy (disambiguation pages), Synonymy (redirect pages) and Associative relationships (hyperlinks between Wikipedia articles). Ciampaglia et al. [2] showed that the distance between two entities within a KG could be used to improve fake news detection accuracy when applying an entity linking method on news articles. In this paper, instead of using the explicit distance between entities, we use the structured relationships constructed by Wikipedia links to analyse the entities within a given sentence using three different methods. Details of these methods are listed below:

2(a). **Similarity of entities**: We follow the method described by Zhu and Iglesias [12]. First, we compute the similarity between two entities using the top 5 concepts with the highest graph-based information content, which are selected and combined into a concept list. The concepts of the Wikipedia KG contain axioms describing concept hierarchies that are usually referred to as ontology classes (type: box), while axioms about entity instances are usually referred as ontology instances (object: a box). Then, we compute the semantic similarity of two entities by calculating the semantic cosine similarity of two concept lists.

2(b). **Relatedness of entities**: We extract the relatedness between two entities using the method described by Witten et. al., [11]

$$sr(a,b) = 1 - \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|)}$$

[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Table 2. Examples of the entity features of each sentence we obtain through entity analysis.

Method name	Aggregation	# of features
Similarity of entities	mean	1
	max	1
Relatedness of entities	mean	1
	max	1
Count of entities	-	1

where a and b are two entities, and A and B are the sets of all the concepts that are linked to a and b. W is the whole set of concepts that appear in all of Wikipedia. $|x|$ is the number of concepts that a given set x contains.

2(c). **Count of entities**: Finally, we also count the non-repeated entities that appear in each sentence, and use the number of the entities as a feature.

As some sentences contain more than two entities, we need to aggregate the similarity and relatedness of each pair of entities into sentence-level features. Therefore, we calculate the mean and max of the similarity and relatedness scores for the pairs of entities within each sentence. Overall, in addition to TF-IDF term features, we therefore have additional 5 features for each sentence, as shown in Table 2.

4 Experimental Setup

In this section, we describe the used dataset, the settings of each component our approach, as well as the evaluation metrics.

Dataset: We use the training and test data provided by the CLEF Check-That! 2019 lab as the training and test datasets, respectively. In the following, we describe in detail the experimental setup we used for the components of our approach:

First-person pronouns resolution and coreference resolution: As mentioned in Section 3, we simply change all the first-person pronouns (i.e., I, we, us, etc) to the current speaker's name. We use Lee et al.'s coreference resolution package[2] to find the entity that a pronoun is referring to. All the parameters are set to their recommended settings [9].

Tokenisation and TF-IDF: We use the Sklearn's TfidfVectorizer[3] to tokenise and calculate the TF-IDF features. All the parameters remain at their default settings.

Entities extraction: In our experiments, we use DBpedia Spotlight[4] to extract entities from each sentence, with the confidence set to 0.3 following [7].

[2] https://github.com/kentonl/e2e-coref

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[4] https://github.com/dbpedia-spotlight/dbpedia-spotlight-model

Table 3. Performances of the top 5 ranked participating groups at the Check-worthiness prediction task.

Team Name	submission	MAP	RR	P@1	P@3	P@5	P@10	P@20	P@50
Copenhagen	primary	0.1660	0.4176	0.2857	0.2381	0.2571	0.2286	0.1571	0.1229
	contr.-1	0.1496	0.3098	0.1429	0.2381	0.2000	0.2000	0.1429	0.1143
	contr.-2	0.1580	0.2740	0.1429	0.1905	0.2286	0.2429	0.1786	0.1200
TheEarthIsFlat	primary	0.1597	0.1953	0.0000	0.0952	0.2286	0.2143	0.1857	0.1457
	contr.-1	0.1453	0.3158	0.2857	0.2381	0.1429	0.1429	0.1357	0.1171
	contr.-2	0.1821	0.4187	0.2857	0.2381	0.2286	0.2286	0.2143	0.1400
IPIPAN	primary	0.1332	0.2865	0.1429	0.0952	0.1430	0.1715	0.1500	0.1171
Terrier	primary	0.1263	0.3254	0.2857	0.2381	0.2000	0.2000	0.1287	0.0915
UAICS	primary	0.1235	0.4650	0.4286	0.2381	0.2286	0.2429	0.1429	0.0944
	contr.-1	0.0649	0.2817	0.1429	0.2381	0.1429	0.1143	0.0786	0.0343
	contr.-2	0.0726	0.4492	0.4286	0.2857	0.1714	0.1143	0.0643	0.0257

Entities analysis: In our experiments, we use the Sematch [13][5]'s KG semantic similarity and relatedness algorithms to calculate the similarity and relatedness of every pair of entities appearing in each sentence. We then calculate the average and maximum similarity scores as well as the relatedness score of a sentence, and use these 4 scores as features. We also count the unique number of entities appearing in each sentence.

Classifier: We tune the SVM classifier's hyperparameters on the training set. In particular, we use the RBF kernel, a C penalty of 10, and a γ of 0.1 in our tuned SVM classifier. Sentences are ranked in descending order by their distance from the classifier's hyperplane.

Evaluation metrics: To evaluate the effectiveness of our approach at highly ranking check-worthy sentences, we use the evaluation metrics suggested by the CheckThat! lab organisers, namely Mean Average Precision (MAP), reciprocal rank (RR), and precision at k (P@k, k={1,3,5,10,20,50}).

5 Results and Discussion

In this section, we address the usefulness of including the pre-processing procedure, as well as the effectiveness of our classification model. In particular, we report and discuss the results of our sentence check-worthiness prediction experiments. Table 3 shows the effectiveness of the top 5 ranked groups that participated in the Lab. Out of a total of 12 groups, our classifier was placed fourth group (ranked by MAP).

Next, to answer whether the pre-processing of data benefits the check-worthiness prediction task, we conduct an ablation study, whereby we remove some of the components of our approach and assessed their resulting performance. Table 4 presents the results using three different variants of our approach: simple TF-IDF model with an SVM classifier, using the pre-processing procedure to

[5] https://github.com/gsi-upm/sematch

Table 4. Performances of our classifier ranking approach with and without pre-processing and entity-based features.

Pre-processing	TFIDF	Entities	MAP	RR	P@1	P@3	P@5	P@10	P@20	P@50
✗	✓	✗	0.0826	0.2000	0.0000	0.0000	0.2000	0.2000	0.3500	0.1571
✓	✓	✗	0.0956	0.2000	0.1667	0.1875	0.1471	0.1587	0.0985	0.0874
✓	✓	✓	0.1263	0.3254	0.2857	0.2381	0.2000	0.2000	0.1287	0.0915

process the debate before using the TF-IDF features and the SVM classifier, and our full approach. The results show that the pre-processing procedure to address pronouns and perform coreference resolution improves MAP performance by 16% (0.0826 → 0.0956). Furthermore, adding the entities features enhances MAP by a further 32% (0.0956 → 0.1263). Thus, we conclude that the pre-processing procedure, as well as our entity-based features are promising and do improve the performance of our approach.

However, as the coreference resolution method cannot achieve a perfect accuracy, the results of our pre-processing procedure are not completely satisfactory. Table 5 shows a clip of one debate, where some sentences' pronouns and coreference resolutions are correct, some are missing, and some are incorrect. Moreover, we do not consider the types of entities in our entity analysis. Such type information may actually be informative, as the entity "*the United States*" may be less informative than "immigration" in an *American* political debate.

6 Conclusions

In this paper, we addressed a task that can be seen as the first step towards fact-checking internet content effectively, as defined by the CLEF CheckThat! 2019 Lab. In particular, we designed a pre-processing procedure, as well as a check-worthiness prediction model, to predict the check-worthiness of each sentence in a given debate. Our experiments showed that the pre-processing procedure, with pronouns resolution and coreference resolution, does improve the performance of the prediction system. Moreover, when using entities extracted and analysed using existing knowledge base tools, the performance of our prediction approach improved further. These findings suggest that pre-processing can be beneficial when analysing text for check-worthiness prediction. They also show that entities analysis might be beneficial in the general fake news detection tasks. In the future, we propose to compare more machine learning methods, and enrich the language processing choices.

Acknowledgements

The first authors acknowledges the support of the China Scholarship Council.

Table 5. An example of the results of the pre-processing procedure. **Bold** denotes the pronouns that should have been changed to the entity it refers to. *Italic* denotes the changed results of the pre-processing procedure. Underline denotes the word is referring to an entity.

Speaker's name	Original text	after pre-processing	type of results
BLITZER	When nearly half of the delegates ..., and the biggest prize of the night is Texas.	When nearly half of the delegates ..., and the biggest prize of the night is Texas .	No entities to be resolved
BLITZER	Immigration is a key issue in **this state**, for all voters nationwide...	Immigration is a key issue in ***Texas*** , for all voters nationwide...	Correct resolution
BLITZER	So, **that**'s where we begin.	So , ***Immigration*** 's where we begin .	Correct resolution
BLITZER	Mr. Trump, you've called for a deportation force to remove the 11 million undocumented immigrants from the United States.	Mr. Trump, you've called for a deportation force to remove the 11 million undocumented immigrants from the United States...	No entities to be resolved
BLITZER	You've also promised to let what you call, "the good ones", come back in.	You 've also promised to let what you call , " the good ones " , come back in .	No entities to be resolved
BLITZER	Your words, "the good ones", after they've been deported.	Your words , " the good ones " , after they 've been deported .	No entities to be resolved
BLITZER	Senator Cruz would not allow them to come back in.	Senator Cruz would not allow ***they*** to come back in .	Incorrect resolution
BLITZER	**He** says that's the biggest difference between the two of you.	***Senator Cruz*** says that 's the biggest difference between the two of you .	Correct resolution
BLITZER	**He** calls your plan amnesty.	He calls the two of you plan amnesty.	Missing resolution

References

1. Atanasova, Pepa and Nakov, Preslav and Karadzhov, Georgi and Mohtarami, Mitra and Da San Martino, Giovanni: Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 1: Check-Worthiness. In Proc. of CLEF-CEUR 2019. (2019)
2. Ciampaglia, Giovanni Luca and Shiralkar, Prashant and Rocha, Luis M and Bollen, Johan and Menczer, Filippo and Flammini, Alessandro: Computational fact checking from knowledge networks. PloS one, 10(6), p.e0128193 (2015)

3. Franco, L Alberto.: Forms of conversation and problem structuring methods: a conceptual development. Journal of the operational research society 57(7), pp. 813–821 (2006)
4. Gencheva, Pepa and Nakov, Preslav and Màrquez, Lluís and Barrón-Cedeño, Alberto and Koychev, Ivan: A context-aware approach for detecting worth-checking claims in political debates. In Proc. of RANLP 2017, pp. 267–276. (2017)
5. Hansen, Casper and Hansen, Christian and Simonsen, J and Lioma, Christina.: The Copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the CLEF-2018 fact checking lab. In Proc. of CLEF-CEUR 2018, pp. 171–175. (2018)
6. Hassan, Naeemul and Zhang, Gensheng and Arslan, Fatma and Caraballo, Josue and Jimenez, Damian and Gawsane, Siddhant and Hasan, Shohedul and Joseph, Minumol and Kulkarni, Aaditya and Nayak, Anil Kumar and other: Claimbuster: The first-ever end-to-end fact-checking system. VLDB Endowment 10(12), pp. 1945–1948 (2017)
7. Joachim Daiber and Max Jakob and Chris Hokamp and Pablo N. Mendes: Improving Efficiency and Accuracy in Multilingual Entity Extraction. In Proc. of I-SEMANTICS 2013. pp. 121-124. (2013)
8. Jaradat, Israa and Gencheva, Pepa and Barrón-Cedeño, Alberto and Màrquez, Lluís and Nakov, Preslav: Claimrank: Detecting check-worthy claims in arabic and english. arXiv preprint arXiv:1804.07587 (2018)
9. Lee, Kenton and He, Luheng and Zettlemoyer, Luke: Higher-order coreference resolution with coarse-to-fine inference. arXiv preprint arXiv:1804.05392. (2018)
10. Patwari, Ayush and Goldwasser, Dan and Bagchi, Saurabh: TATHYA: A multi-classifier system for detecting check-worthy statements in political debates. In Proc. of CIKM 2017, pp. 2259–2262. (2017)
11. Witten, Ian H and Milne, David : An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proc. of WAI workshop at AAAI 2008, pp. 25–30 (2008)
12. Zhu, Ganggao and Iglesias, Carlos Angel: Computing semantic similarity of concepts in knowledge graphs. IEEE Transactions on Knowledge and Data Engineering 29(1), pp. 72–85 (2016)
13. Zhu, Ganggao and Iglesias Fernandez, Carlos Angel: Sematch: semantic entity search from knowledge graph. Telecomunicacion, pp. 1–12 (2015)
14. Zuo, Chaoyuan and Karakas, Ayla Ida and Banerjee, Ritwik: A hybrid recognition system for check-worthy claims using heuristics and supervised learning. In Proc. of CLEF-CEUR 2018, pp. 171–175. (2018)