Team GPLSI. Approach for automated fact checking

Aimée Alonso-Reina, Robiert Sepúlveda-Torres, Estela Saquete*, Manuel Palomar*
University of Havana
I.U for Computer Research, Univesity of Alicant
* Department of Software and Computing systems, Univesity of Alicant
aimeear1993@gmail.com, rsepulveda911112@gmail.com, {stela, mpalomar}@dlsi.ua.es

Abstract

Fever Shared 2.0 Task is a challenge meant for developing automated fact checking systems. Our approach for the Fever 2.0 is based on a previous proposal developed by Team Athene UKP TU Darmstadt. Our proposal modifies the sentence retrieval phase, using statement extraction and representation in the form of triplets (subject, object, action). Triplets are extracted from the claim and compare to triplets extracted from Wikipedia articles using semantic similarity. Our results are satisfactory but there is room for improvement.

1 Introduction

The proliferation of user-generated content and Computer Mediated Communication (CMC) technologies, such as blogs, Twitter, and other social media enable mass scale news delivery mechanisms (Conroy, Rubin, & Chen, 2015; Rubin & Lukoianova, 2015). The emergence of social networks and their use for the dissemination of news are a double-edged sword. On the one hand, its low cost, easy access and rapid distribution of information encourages people to search and consume news from social networks. On the other hand, it allows the proliferation of "fake news", i.e., low quality news with intentionally false information (Shu, Sliva, Wang, Tang, & Liu, 2017).

Automated fact checking for proving news veracity by reliable sources is a vital task related to the processes of fake news detection (Bondielli & Marcelloni, 2019). It consists of classifying the veracity of each news item by assigning a veracity value. Artificial Intelligence (AI) techniques are applied in order to automate this process. Computational fact checking may significantly enhance our ability to evaluate the veracity of dubious information (Ciampaglia et al., 2015). The work of (Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018) has resulted in the development of a dataset containing facts with their corresponding classification, and evidences. This was applied in Fever 1.0 Shared Task.

The dataset was obtained by generating claims and recovering their corresponding evidence from Wikipedia. This crowd-sourced online encyclopedia has been shown to be nearly as reliable as traditional encyclopedias, despite covering many more topics (Ciampaglia et al., 2015).

Fever Shared Task is a challenge meant for developing automated fact checking systems. The central component is a trained dataset for creating new models, applying AI techniques to recognize patterns contained in the dataset. An example of a claim, evidence and classification tuple is shown in figure 1.

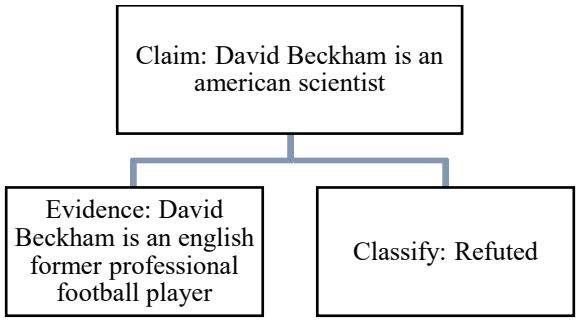

Figure 1: Fever dataset. Tuple example.

Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 110–114
Hong Kong, November 3, 2019.

Finally, there is a testing strategy that automatically classifies claims and generates evidences.

We are proposing a modified approach to the Fever 2.0 Shared Task by adapting a previous proposal developed by Team Athene UKP TU Darmstadt (Hanselowski et al., 2018).

We agreed with the Fever baseline and the Team Athene and divided the process into three tasks: document retrieval; sentence retrieval; and, recognizing textual entailment.

A non-formal diagram illustrates the relations among the tasks that are applied in our proposal. The shadowed frame shows the task that we have modified (see figure 2).

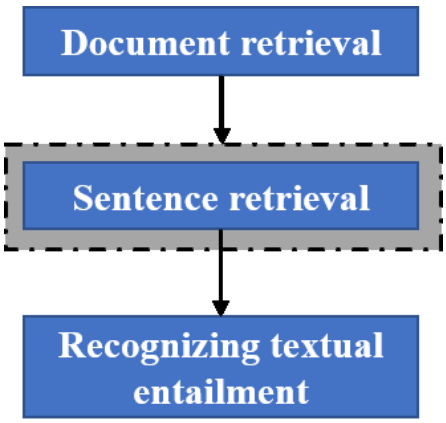

Figure 2: System internal structure.

2 Document retrieval

The main goal of the document retrieval task is to obtain relevant pages, using Wikipedia as a data source. This task retrieves those pages containing elements related to the claim under evaluation.

For each claim, a set of noun phrases is extracted and used for indexing the pages containing these terms.

The library AllenNLP (Gardner et al., 2018) is applied for extracting the noun phrases for each claim and a Wikipedia proprietary library is used for indexing.

Three alternatives for document retrieval were considered. First, the baseline proposal (Thorne et al., 2018), a basic approach that might present information loss. Second, Apache Lucene, although is a robust tool, its integration with our approach was very complex and inefficient. Third, the Team Athene proposal, was considered by our team as the best option for this task because it combines accuracy with simplicity.

3 Sentence retrieval

To select the sentences best related to the claim under analysis, a sentence retrieval task is defined. In this step, the Team Athene approach was selecting candidate sentences as a potential evidence set for a claim. These sentences were extracted from the Wikipedia articles retrieved during the document retrieval phase (Hanselowski et al., 2018).

Our approach uses statement extraction and representation in the form of triplets (subject, object, action) to represent the information transmitted by a sentence. These triplets are extracted from the claim using the statement detector defined in the paper (Estevez-Velarde et al., 2018).

Through triplet comparison, we aim to determine whether the facts from the claim are supported by the information source. To compare triplets, we used *Spacy*'s model *en_core_web_lg* (Honnibal & Montani, 2017), which is one of the new neuronal models of SpaCy v2.0 for labeling, analysis and entity recognition.

The semantic similarity between two texts can be defined as $S: S \in \mathbb{R}$; $S \in [0,1]$. When two triplets are compared, three semantic similarities are extracted: similarity between subjects (s_s), similarity between objects (s_o), and similarity between actions (s_A). To decide which triplets are more similar, we use the average $A = avg(s_s, s_o, s_A)$ and the minimum of these three similarities $M = min(s_s, s_o, s_A)$. Two triplets are more similar, the bigger their minimum similarity is, and as a tie-breaker for minimum similarities, the average is used.

To select the facts that best match the claim, a score was defined, combining the Team Athene score with our similarity measures. At testing time, Team Athene calculated a score between a claim and each sentence from the retrieved documents. With that purpose, an ensemble of ten models with different random seeds was deployed (Hanselowski et al., 2018). They calculated the mean score of a claim-sentence pair over all ten models of the ensemble and established a ranking for all pairs. Finally, the sentences with the highest-ranked pairs were provided as an output of the model.

Defining n as the number of sentences to be extracted, the first step of our ranking algorithm is to select the $n * 3$ sentences best ranked by the

Team Athene score. The next step is to extract (subject, object, action) triplets for each sentence and compare these triplets to the ones extracted from the claim. Finally, we sort the sentences according to their similarity with the claim triplets, obtaining as evidence n sentences, (those considered more similar to the claim).

When extracting statements from a sentence, it is important to note whether the sentence can be considered a negation. This is implemented by checking for keywords such as *"never", "not"*, among others. Initially, when comparing two triplets, if only one of them was negated, they were considered not similar (based on a semantic perspective). Hence, the action similarity (s_A) was set to 0 and these triplets would not be taken to account by the ranking algorithm. However, after receiving feedback on the testing results, we realized that this approach was affecting the retrieval of evidence that refutes the claim. In order to provide a solution for this issue, we removed the restriction that was keeping these negated sentences of being taken to account by the ranking algorithm, creating an opportunity for them to be highly scored in the sentence retrieval phase and used as evidence.

In comparison with the Team Athene proposal, this phase contains a significant change that might vary the final results for the next phase.

4 Recognizing textual entailment

This task classifies the claims versus the supposed evidences that are obtained from previous tasks. It is well known as an active research area in Natural Language Processing in the last decade. That is corroborated by the number of related papers (Korman, Mack, Jett, & Renear, 2018; Padó, Noh, Stern, Wang, & Zanoli, 2015; Paria, Annervaz, Dukkipati, Chatterjee, & Podder, 2016).

A description for Stanford Natural Language Inference (SNLI) dataset is reported in (Bowman, Angeli, Potts, & Manning, 2015) and the development of multi-Genre Natural Language Inference (MultiNLI) may be consulted at (Williams, Nangia, & Bowman, 2017). Both of them were applied for training complex NLI models.

The Enhanced Sequential Inference Model (ESIM) (Chen et al., 2016) is one of the most commonly applied for accomplishing the recognizing textual entailment task . This model has been trained over different proposals by applying minimal changes into neural network parameters.

The ESIM model extended by (Hanselowski et al., 2018) is the one used in our proposal. The input is a set of ordered pairs, composed of the same claim and five sentences selected from the previous tasks.

Each word from these pairs is represented as a vector by concatenating two word embeddings. In this case, FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) and Glove (Pennington, Socher, & Manning, 2014) are applied. These word embeddings are selected because they have been previously trained with Wikipedia information. The vectors are passed to the model for the training and testing phases.

5 Results

Our results differ discretely from the Team Athene proposal. A more extensive experimentation is recommended in order to improve the final claim classification in comparison to that obtained by Athene.

The results obtained from the document retrieval task are coincident with the original proposal (Hanselowski et al., 2018), because the model applied for obtaining Wikipedia pages is the same.

To accomplish the sentence retrieval task, five sentences were selected, ranked according to the best score.

To show the differences between the sentence retrieval task of our approach and that of the Team Athene, all the evidences sets are collected and compared in this task. We used the “Shared Task Development Dataset (Labelled)”. This dataset contains 19,998 tuples equal to that of the “Shared Task Blind Test Dataset (Unlabelled)” which was used to submit our predictions. Table 1 shows the comparative result.

Variation in the evidence sets	Count	%
Nil	7988	39.94
One variation	10582	52.91
Two variations	1265	6.32
Three variations	119	0.59
Four variations	30	0.15
Five variations	14	0.07

Table 1. Result of evidence sets comparison.

As can be seen in Table 1, an intuitive analysis was carried out that allows us to believe that the results of the retrieval sentence task are different between the teams. This implies changes in the result classification for the textual entailment task.

Moreover, we calculate the accuracy of the collected evidence sets for the "Shared Task Development Dataset (Labelled)". These results are shown in table 2 for Team Athene and table 3 for our team.

Evidence sets	**Team Athene**	**%**
At least one evidence in common	12594	62.97
All evidences are different	7404	37.02

Table 2. The Team Athene accuracy in terms of finding one evidence in common.

Evidence sets	**Team GPLSI**	**%**
At least one evidence in common	12547	62.74
All evidence different	7451	37.25

Table 3. The Team GPLSI accuracy in terms of finding one evidence in common.

The accuracy of sentence retrieval for two teams is similar. This low score affects negatively on the calculation of the Fever score.

The final task of our proposal aims to classify the claim as one of three classes: "SUPPORTS", "REFUTES", "NOT ENOUGH INFO". This task does not differ from the Team Athene approach. However, expected differences among the results are obtained, albeit with low percentage between teams. The changes proposed for the sentence retrieval task and the differences among sentences justify these results.

Table 4 shows the results from participant teams on Fever 2.0 Shared Task, the three best teams from last year (2018), and Fever Baseline. The results are ordered considering the Fever Score for each team.

Team	**Resilience (%)**	**Fever Score (%)**
Dominiks	35.82	68.46
CUNLP	32.92	67.08
UNC	30.47	64.21
UCL MR	35.82	62.52
Athene	25.35	61.58
GPLSI	19.63	58.07
CalcWorks	DNQ	33.56
Baseline	11.06	27.45

Table 4. Main results of the challenge.

The updated code of our approach may be accessed at URL: https://github.com/rsepulveda911112/fever-2019-team-gplsi

6 Conclusions

The GPLSI team has developed an automated system that modifies the sentence retrieval task drastically and get similar results. The relevance of the applied model for obtaining triplets and similarity metrics are confirmed.

We consider that to improve the fever score we must improve the accuracy of the sentence retrieval task.

For the task of recognizing textual entailment in the future, we think that the classification can be improved by incorporating features into the ESIM model. These characteristics should improve both the detection of contradictions that would deliver the classification "REFUTES" and, the accuracy of the "NOT ENGOUGH INFO" classification when there is a lack of relevant data that can refute or support a claim.

Acknowledgments

This research work has been partially funded by the University of Alicante (Spain), Generalitat Valenciana and the Spanish Government through the projects Tecnologías del Lenguaje Humano para una Sociedad Inclusiva Igualitaria y Accesible (PROMETEU/2018/089), Modelado del Comportamiento de Entidades Digitales mediante Tecnologías del Lenguaje Humano (RTI2018-094653-B-C22) and Integer: Intelligent Text Generation, Generación Inteligente de Textos (RTI2018-094649-B-I00).

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55. https://doi.org/10.1016/j.ins.2019.05.035

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. Retrieved from http://arxiv.org/abs/1508.05326

Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., &

Inkpen, D. (2016). Enhanced LSTM for Natural Language Inference. Retrieved from http://arxiv.org/abs/1609.06038

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational Fact Checking from Knowledge Networks. *PLOS ONE*, *10*(6), e0128193. https://doi.org/10.1371/journal.pone.0128193

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* (p. 82). American Society for Information Science.

Estevez-Velarde, S., Gutierrez, Y., Montoyo, A., Piad-Morffis, A., Munoz, R., & Almeida-Cruz, Y. (2018). Gathering object interactions as semantic knowledge. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)* (pp. 363–369). The Steering Committee of The World Congress in Computer Science, Computer ….

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., … Zettlemoyer, L. (2018). AllenNLP: A Deep Semantic Natural Language Processing Platform. Retrieved from http://arxiv.org/abs/1803.07640

Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., & Gurevych, I. (2018). *UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification*. Retrieved from https://www.ukp.tu-darmstadt.de/

Honnibal, M., & Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To Appear*, *7*.

Korman, D. Z., Mack, E., Jett, J., & Renear, A. H. (2018). Defining textual entailment. *Journal of the Association for Information Science and Technology*, *69*(6), 763–772. https://doi.org/10.1002/asi.24007

Padó, S., Noh, T.-G., Stern, A., Wang, R., & Zanoli, R. (2015). Design and realization of a modular architecture for textual entailment. *Natural Language Engineering*, *21*(2), 167–200. https://doi.org/10.1017/s1351324913000351

Paria, B., Annervaz, K. M., Dukkipati, A., Chatterjee, A., & Podder, S. (2016). A Neural Architecture Mimicking Humans End-to-End for Natural Language Inference. Retrieved from http://arxiv.org/abs/1611.04741

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. Retrieved from http://nlp.

Rubin, V. L., & Lukoianova, T. (2015). Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, *66*(5), 905–917.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and verification. *ArXiv Preprint ArXiv:1803.05355*.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. Retrieved from http://arxiv.org/abs/1704.05426