UPV-UMA at CheckThat! Lab: Verifying Arabic Claims using a Cross Lingual Approach

Bilal Ghanem[1], Goran Glavaš[2], Anastasia Giachanou[1], Simone Paolo Ponzetto[2], Paolo Rosso[1], and Francisco Rangel[1,3]

[1] PRHLT Research Center, Universitat Politècnica de València, Spain
{bigha@doctor., angia9@, prosso@dsic.}upv.es
[2] University of Mannheim, Germany
{goran, simone}@informatik.uni-mannheim.de
[3] Autoritas Consulting, Valencia, Spain
francisco.rangel@autoritas.es

Abstract. In this paper we present our team participation at CheckThat!-2019 lab - Task 2 on Arabic claim verification. We propose a cross-lingual approach to detect the factuality of claims using three main steps, evidence retrieval, evidence ranking, and textual entailment. Our approach achieves the best performance in subtask-D, with a value of 0.62 as F1.

Keywords: Claims Factuality · Arabic · Evidence Retrieval · Cross-Lingual Word Embeddings

1 Introduction

Rumours in news media and political debates may shape people's believes. Public opinion can be easily manipulated and this sometimes can lead to severe consequences including harming individuals, religions, and several other victims. For example, in 2016 a man opened fire on a Washington pizzeria because of a fake claim that reported that the pizzeria was housing young children as sex slaves as part of a child abuse ring led by the presidential candidate Hillary Clinton [16]. The spread of these claims is rapid and uncontrolled, which makes their verification hard and time consuming. Thus, automated methods have been proposed to facilitate the process of their verification.

The Arabic language has a large number of speakers around the world. However, due to the language has a limited number of Natural Language Processing (NLP) resources for the Arabic language, there is an increasing gap between this language and other languages regarding the availability of NLP systems. Recently, there have been various research attempts on NLP tasks on Arabic, such as fact checking [12] [4], author profiling [14] [13], and irony detection [9].

In this paper, we present our participation in the CheckThat! Lab - Task 2 [7] for detecting the factuality of Arabic claims in general news topics. Our approach

is based on inferring the veracity by using a Natural Language Inference (NLI) system trained on the English language to predict if an Arabic pair of sentences entail each other. To do that, we use cross-lingual embeddings.

2 Related Work

Previous works on claims' factuality can be roughly split into two main approaches: external sources-based, and context-based. The external-sources-based approaches pass a claim to external search engines (e.g., Google, Bing), and then they build various features from the results. Ghanem et al. [5] proposed to pass the claims to Google and Bing search engines in order to retrieve evidences and then they extracted features like similarity between the claims and the snippets, as well as the Alexa rank[4] of the retrieved links. Finally, the authors used these features to train a Random Forest classifier. A similar approach was proposed by Karadzhov et al. [8] who computed the cosine similarity between the claim and the top N results to feed these similarities into a Long-Short Term Memory (LSTM).

On the other hand, the context-based approaches use a different way of inferring the factuality. Castillo et al. [1] used text characteristics, user-based, topic-based, and tweets propagation-based features. Similarly, Mukherjee and Weikum [11] proposed a continuous conditional random field model that exploits several signals of interaction between a set of features (e.g., language of the news, source trustworthiness, and users' confidence).

3 Task Description

Given a set of Arabic claims with their relevant documents (web pages), the goal of the task is to predict the factuality of these claims using the provided web pages. Task 2[5] has 4 different sub-tasks, but we decided to participate in two of them, namely task B and D. Task B aims to predict how useful is a web page with respect to a claim, and the target labels are: *very useful for verification*, *useful for verification*, *not useful* or *not relevant*. Task D aims to find the claim's factuality (*True* or *False*). This task is organized in 2 cycles; in cycle 1 the factuality should be estimated using the provided unlabeled web pages, whereas in cycle 2 using useful web pages (very useful and useful labels). The organizers provided the web pages in a real scenario, where the participants had to retrieve the evidence and then compared it to the claim.

Regarding the task data, the organizers provided 10 Arabic claims with their correspondent web pages with a number between 26 and 50 web pages results for each claim. These web pages were provided in their original form ($HTML$ format). For the test set, the organizers provided 59 claims to be verified.

[4] https://www.alexa.com/siteinfo
[5] https://sites.google.com/view/clef2019-checkthat/task-2-evidence-factuality

4 Proposed Approach

We propose an approach that consists of the following three main steps: evidence retrieval, evidence ranking, and textual entailment. Figure 1 shows a schematic overview of our approach.

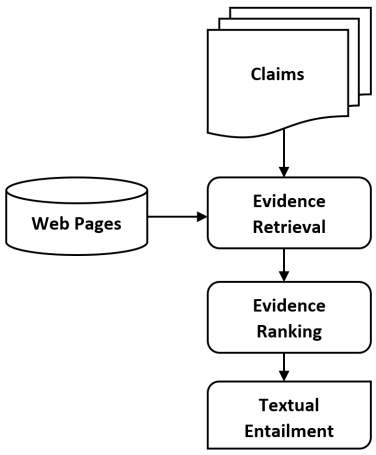


Fig. 1: Overview of our approach.

Evidence Retrieval: In the first step, we read the content of the articles and then we split them into sentences using *comma* (,) and *dot* (.) as delimiters following the previous literature work [10]. To obtain the best recall, we retrieve the top N similar sentences to the claim using cosine similarity over character n-grams. We use n-gram of length 5 and 6; we choose them experimentally. In addition, we tried to retrieve the most similar sentences using Named Entities (NEs), but we found that there are some sentences without named entities, like:

تخفف المشروبات الساخنة من نزلات البرد وتقلل من أعراضها

Translation: *Cold drinks reduce colds and their symptoms*

In this step, we discard very short sentences[6]. Finally, we pass the top 20 sentences to the next step.

Evidence Ranking: For this step, we rank the top 20 sentences using word embeddings. For each claim-evidence pair, we measure their similarity and we rank the evidence based on the similarity values. For the word embeddings, we

[6] We discarded sentences that have less than 35 characters. This kind of sentences appeared when a dot and a comma occur closely.

use Arabic $fastText$[7] pretrained model. We explore the following three different similarity techniques:

1. *Cosine over embeddings:* We calculate the average of the words embeddings of each sentence, and compute the cosine similarity.
2. *Cosine over weighted embeddings:* We calculate the average of the words' embeddings weighted by the Term Frequency Inverse Document Frequency (TF-IDF) weighting scheme, and then we compute the cosine similarity on the two weighted sentences' vectors. We compute the TF-IDF weights using the Comparable Wikipedia Corpus [15].
3. *DynaMax:* It is an unsupervised and non-parametric similarity measure based on fuzzy theory that dynamically extracts good features from the word embeddings depending on the sentence pair [19].

Since the training dataset is very small, it was not possible to find the best similarity technique statistically. Thus, we decided to manually investigate the ranked sentences and we found that using *DynaMax* we get the most semantically similar evidence sentences at the top ranks.

Textual Entailment: For this step, we propose to train a system on par with state-of-the-art results in NLI task, that is the Enhanced Sequential Inference Model (ESIM) [2]. We follow the implementation details of [18]. We train the ESIM on a large NLI corpus for English, namely MultiNLI [18]. Since the claims' language is Arabic, we first project the Arabic word embeddings to the vectors space of the English word embeddings[8] we used during the training of the ESIM model. To this end, we learn a linear projection matrix by solving the Procrustes problem [17,6] using 5K automatically obtained English-Arabic word translations as supervision[9]. To evaluate the performance of our model, we use a multilingual XNLI corpus [3] created by translating development and test sets of the MultiNLI corpus. Our cross-lingually transferred ESIM system achieved 58% accuracy on the Arabic test set of the XNLI corpus.

In this step of our approach, we receive a claim with its 20 ranked sentences from the *Evidence Ranking* step. We feed the claim with each ranked sentence to the ESIM model and we estimate their prediction probabilities with respect to Entailment, Neutral, Contradiction labels. Since each claim is represented by 20 predictions, we weight the predictions in one of two methods:

1. **Similarity Weighting**: We weight the predictions by the evidence ranking similarity values. Given the prediction probability P of one of the classes C, we weight it as: $P_c = \sum_{i=1}^{20} P_{ci} * SentenceSimilarity_i$.
2. **Majority Class**: Given the NLI predictions for each claim P, we extract the majority class by: $count_{classes}(\ argmax\ P)$.

[7] https://fasttext.cc/docs/en/crawl-vectors.html

[8] We used English fastText embeddings: https://github.com/facebookresearch/fastText

[9] The 5k words obtained by translating the most frequent words appeared in an English Wikipedia corpora using Google Translator.

Finally, after weighting the predictions for each claim, we infer the final 2-classes prediction (True, False) from the 3-classes (NLI labels) using the following rule:

$$f(P_{entailment}, P_{contradiction}) = \begin{cases} True, & \text{if } P_{entailment} \geq P_{contradiction} \\ False, & \text{otherwise} \end{cases}$$

For the Majority Class weighing method, the $P_{entailment}$ and $P_{contradiction}$ of a claim are represented by the count frequency of the class instead of its probability.

5 Experiments and Results

Task2 subtask-B: In this subtask, we use the first two steps of our approach to submit a run. In the first step, we retrieve the sentences from the web pages using character n-grams. Here, we retrieve all the sentences with a cosine similarity value greater than 0. Then, we pass them to the next step where we rank them based on the words embeddings. At this step, we discard the ranks and we only average the sentences similarity values for each web page (WP_{avg}). Then with a rule-based method, we map the web pages averaged values into the 4 classes:

$$f(WP_{avg}) = \begin{cases} very\_useful, & \text{if } WP_{avg} \geq 0.45 \\ useful, & \text{if } WP_{avg} > 0.35 \ \& \ WP_{avg} < 0.45 \\ not\_useful, & \text{if } WP_{avg} \leq 0.35 \\ not\_relevant, & \text{if } WP_{avg} = -1 \end{cases}$$

In the cases that we do not get any sentence from the retrieval process, we set WP_{avg} to -1. The thresholds are set experimentally. Table 1 presents the results of the subtask-B, for both 2-classes and 4-classes prediction. Our submission for the 2-classes prediction obtains the best performance, but still lower than the provided baseline by the organizers. For the 4-classes prediction, we obtain a lower overall rank, lower than the baseline as well.

Task2 subtask-D: For subtask-D, we use our three steps approach. For each of the two cycles (see Section on Task Description) we submit two runs, one using the *Similarity Weighting* and the other using the *Majority Class*[10].

Table 2 presents the results on the test set for the subtask-D. Considering the second cycle submissions' results, since they are less biased, we observe that the similarity value weighting performs better than the majority class method clearly. We obtain the best performing runs in both cycles, higher than the baselines with 0.25 F1 value on average.

[10] We submitted our runs for cycle-1 at late time, thus the organizers considered them as submissions for cycle-2.

Table 1: The subtask-B results in terms of Accuracy, Precision, Recall, and F1 metrics.

Evaluation criteria	Acc.	Prec.	Recall	F1
2-classes prediction				
Baseline	0.57	0.30	0.72	0.42
2-classes submission	0.49	0.26	0.73	0.38
4-classes prediction				
Baseline	0.30	0.32	0.32	0.28
4-classes submission	0.24	0.3	0.29	0.23

Table 2: The subtask-D results in terms of Accuracy, Precision, Recall, and F1 metrics.

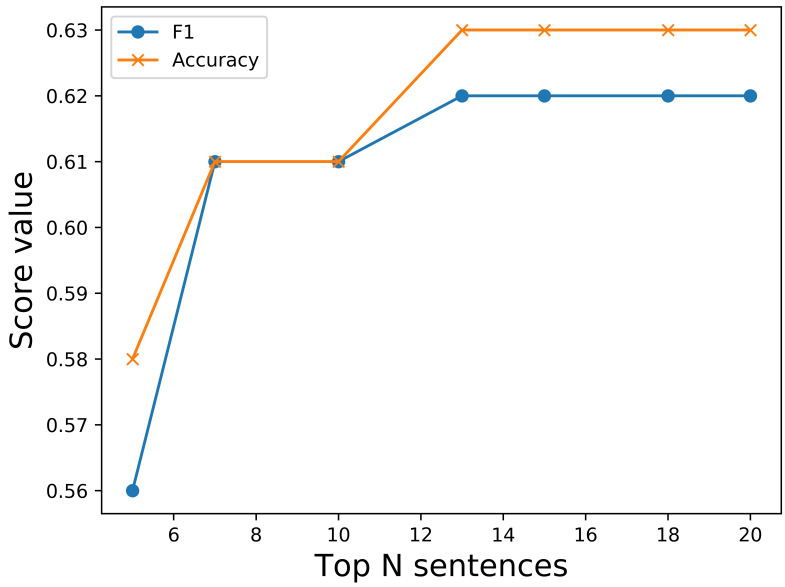
run #	method	Acc.	Prec.	Recall	F1
Cycle-1: Unlabeled web pages					
1	Similarity Value	0.56	0.56	0.56	0.55
2	Majority Class	0.58	0.65	0.57	0.51
-	Baseline	0.51	0.25	0.50	0.34
Cycle-2: Useful web pages					
1	Similarity Value	0.63	0.63	0.63	0.62
2	Majority Class	0.58	0.60	0.57	0.54
-	Baseline	0.51	0.25	0.50	0.34

6 Analysis

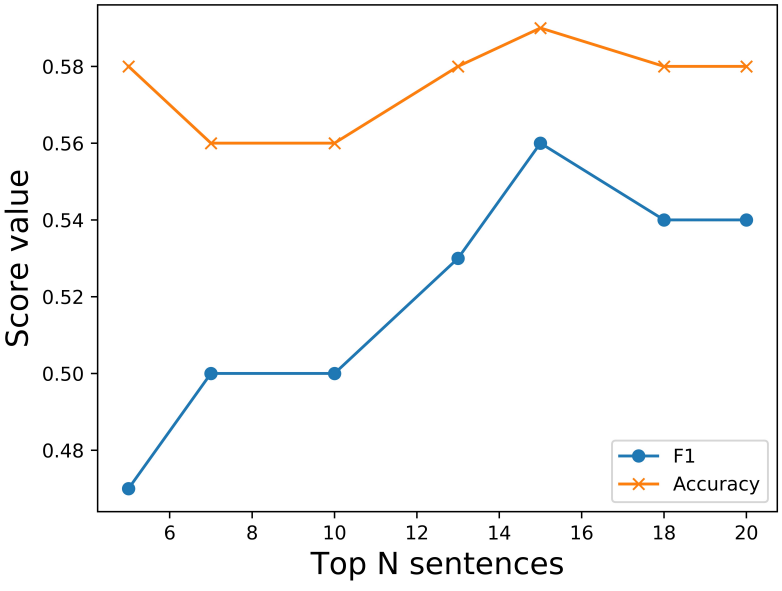
In our experiments we consider the first 20 sentences to be fed to the ESIM model. In Figure 2, we investigate the effect of varying the number of sentences to consider for each claim on the test set. We use the second cycle (given the labeled web pages) in this experiment.

Understanding the causes of errors of our approach is important for future improvements. We manually examined the predictions to understand the causes of errors. We categorize them into the following cases:

1. **Un-famous news**: Some of the truthful claims were not covered by many news sites. We found that our approach retrieved few correct evidence (two or three evidence) while the rest of the evidence describe things related to the main entity but not regarding the same claim issue. Since in our approach we use the first 20 evidence to infer the factuality, the first 3 similar evidence, as an example, voted positively for the factuality of the sentence, where the rest 17 voted negatively. This kind of errors can be resolved by using a dynamic number of evidence sentences for each claim instead of a fixed one.
2. **The spread of false rumors**: The spread of rumors over the web can mislead people. Since our approach is based on retrieving the claim's evidence from the web, the existence of these false rumors can consequently mislead our system. As an example, given the following false claim:



(a)



(b)

Fig. 2: The performance of our approach on the test set using (**a**) the Similarity Value weighting and (**b**) Majority Class with varying the number of evidence sentences.

توفي رفعت الأسد عم بشار الأسد في أحد مستشفيات باريس

Translation: *Rifaat al-Assad, the uncle of Bashar al-Assad, died in a hospital in Paris*

Our approach retrieved the following evidence which supports the claims:

أنباء عن وفاة جزار حماة وسجن تدمر؛ رفعت الأسد في أحد مستشفيات باريس

Translation: *News about the death of the butcher of Hama and Palmyra prisons, Rifaat al-Assad in a Paris hospital*

This evidence was retrieved as a Twitter post. Considering only news agencies as source of news where random users are not allowed to post news, can prevent these errors.

3. **Inaccurate sentence segmentation**: The Arabic language has a complicated sentence structure, where using dots to split a document into sentences is inaccurate step. Following the previous works in Arabic, we used *dot* (.) and *comma* (,) to split the evidence documents into sentences. We found that in some cases, the important evidence sentence in a document has a *comma* between the object and predicate. As an example:

أعدمت مصر ١٥ متشددا أدينوا بشن هجمات نتج عنها
مقتل عدد من رجال الجيش والشرطة في شبه جزيرة سيناء

Translation: *Egypt executed 15 militants convicted of attacks that resulted in the deaths of a number of military and police men in the Sinai Peninsula*

The evidence in a document presented as follows:

نفذت مصلحة السجون رابع حكم بالإعدام فى ١٥ متهماً،
على خلفية اتهامهم بقتل ضباط وجنود القوات المسلحة فى شمال سيناء

Translation: *The Prison Service carried out a fourth death sentence in 15 accused, (**COMMA**) for killing officers and soldiers of the armed forces in northern Sinai*

The *comma* between the sentence's parts made the evidence unsupportive to the claim by splitting it.

4. **Weak ESIM predictions**: We found some claims whose evidence was retrieved correctly but the ESIM model was unable to verify them. We argue that this kind of error is due to the aligned cross-lingual embedding.

7 Conclusion and Future Work

In this paper, we presented our participation in CheckThat! lab - Task 2 at CLEF-2019. We presented an approach that consists of 3 main steps from Arabic claims verification. Our proposed approach managed to achieve a good performance. Also, from the error analysis, the results showed that our cross-lingual model is solid since the majority of errors were due to the other previous reasons. As a future work, we plan to focus and improve the errors cases we identified for more effective retrieval, ranking, and prediction.

Acknowledgements

The work of Paolo Rosso and Fracisco Rangel was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. The work of Paolo Rosso was partially funded by the the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). The work of Goran Glavaš was carried out within the scope of the AGREE project supported by the Eliteprogramm of the Baden-Wrttemberg Stiftung. Anastasia Giachanou is supported by the SNSF Early Postdoc Mobility grant P2TIP2_181441 under the project Early Fake News Detection on Social Media, Switzerland

References

1. Castillo, C., Mendoza, M., Poblete, B.: Information Credibility on Twitter. In: Proceedings of the 20th international conference on World Wide Web. pp. 675–684 (2011)
2. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for Natural Language Inference. arXiv preprint arXiv:1609.06038 (2016)
3. Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: Xnli: Evaluating Cross-lingual Sentence Representations. arXiv preprint arXiv:1809.05053 (2018)
4. Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Da San Martino, G., Atanasova, P.: Overview of the CLEF-2019 CheckThat!: Automatic Identification and Verification of Claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. LNCS, Lugano, Switzerland (September 2019)
5. Ghanem, B., Montes-y Gómez, M., Rangel, F., Rosso, P.: UPV-INAOE-Autoritas-Check That: An Approach based on External Sources to Detect Claims Credibility. Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, CLEF '18, Avignon, France, September. (2018)
6. Glavas, G., Litschko, R., Ruder, S., Vulic, I.: How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. arXiv preprint arXiv:1902.00508 (2019)

7. Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeño, A., Nakov, P.: Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. CEUR Workshop Proceedings, CEUR-WS.org, Lugano, Switzerland (2019)
8. Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I.: Fully Automated Fact Checking Using External Sources. arXiv preprint arXiv:1710.00341 (2017)
9. Karoui, J., Zitoune, F.B., Moriceau, V.: Soukhria: Towards an Irony Detection System for Arabic in Social Media. Procedia Computer Science **117**, 161–168 (2017)
10. Lee, Y.S., Papineni, K., Roukos, S., Emam, O., Hassan, H.: Language Model Based Arabic Word Segmentation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. pp. 399–406. Association for Computational Linguistics (2003)
11. Mukherjee, S., Weikum, G.: Leveraging Joint Interactions for Credibility Analysis in News Communities. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 353–362 (2015)
12. Nakov, P., Barrón-Cedeno, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 372–387 (2018)
13. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN17. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 275–290. Springer (2017)
14. Rosso, P., Rangel Pardo, F.M., Ghanem, B., Charfi, A.: ARAP: Arabic Author Profiling Project for Cyber-Security. Sociedad Española para el Procesamiento del Lenguaje Natural (2018)
15. Saad, M., Alijla, B.O.: Wikidocsaligner: An off-the-shelf Wikipedia Documents Alignment Tool. In: Proceedings of the 2017 Palestinian International Conference on Information and Communication Technology. pp. 34–39 (2017)
16. Simpson, I.: Man pleads guilty in washington pizzeria shooting over fake news. https://www.reuters.com/article/us-washingtondc-gunman/man-pleads-guilty-in-washington-pizzeria-shooting-over-fake-news-idUSKBN16V1XC (2017), [Online; accessed 10-may-2019]
17. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. In: Proceedings of ICLR (2017), https://arxiv.org/abs/1702.03859
18. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage Challenge Corpus for Sentence Understanding Through Inference. arXiv preprint arXiv:1704.05426 (2017)
19. Zhelezniak, V., Savkov, A., Shen, A., Moramarco, F., Flann, J., Hammerla, N.Y.: Don't Settle for Average, Go for the Max: Fuzzy Sets and Max-Pooled Word Vectors. arXiv preprint arXiv:1904.13264 (2019)