Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification.

Anonymous EMNLP submission

Abstract

This paper presents our submission for the FEVER Workshop Shared Task. Our system is an end-to-end pipeline that extracts factual evidence from Wikipedia and infers a decision about the truthfulness of the claim based on the extracted evidence. On the development set[1] of the FEVER Shared task (Thorne et al., 2018), our pipeline achieves significant improvement over the baseline for all the components (Document Retrieval, Sentence Selection and Textual Entailment), with a final FEVER score of 50.83 compared to 31.27 of the baseline system.

1 Introduction and Background

Fact checking is a type of investigative journalism where experts examine the claims published by others for their veracity. The claims can range from statements made by public figures to stories reported by other publishers. The end goal of a fact checking system is to provide a verdict on whether the claim is true, false, or mixed. Several organizations such as `FactCheck.org` and `PolitiFact` are devoted to such activities.

The FEVER Shared task aims to evaluate the ability of a system to verify information using evidence from Wikipedia. Given a claim involving one or more entities (mapping to Wikipedia pages), the system must extract textual evidence (sets of sentences from Wikipedia pages) that support or refute the claim and then using this evidence, label the claim as Supported, Refuted or NotEnoughInfo. The dataset for the shared task was introduced by Thorne et al. (2018) and consists of 185,445 claims. Table 1 shows three instances from the data set with the claim, the evidence and the verdict.

Claim : Fox 2000 Pictures released the film Soul Food. **[wiki/Soul_Food_(film)]** **Evidence**: Soul Food is a 1997 American comedy-drama film produced by Kenneth ”Babyface” Edmonds , Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures . **Verdict**: **SUPPORTS**
Claim : Murda Beatz’s real name is Marshall Mathers. **[wiki/Murda_Beatz]** **Evidence**: Shane Lee Lindstrom (born February 11, 1994), known professionally as Murda Beatz, is a Canadian hip hop record producer and songwriter from Fort Erie, Ontario. **Verdict**: **REFUTES**
Claim : L.A. Reid has served as the CEO of Arista Records for four years. **[wiki/L.A._Reid]** **Evidence**: He has served as the chairman and CEO of Epic Records, a division of Sony Music Entertainment, the president and CEO of Arista Records, and the chairman and CEO of the Island Def Jam Music Group. **Verdict**: **NOT ENOUGH INFO**

Table 1: Examples of claims, the extracted evidence from Wikipedia and the verdicts from the shared task dataset (Thorne et al., 2018)

The baseline system described by Thorne et al.(2018) uses 3 major components:

- **Document Retrieval**: Given a claim, identify relevant documents from Wikipedia which contain the evidence to verify the claim. Thorne et al.(2018) used the document retrieval component from the DrQA system (Chen et al., 2017), which returns the k nearest documents for a query using cosine similarity between binned unigram and bigram TF-IDF vectors.

- **Sentence Selection**: Given the set of retrieved document, identify the candidate evidence sentences. Thorne et al.(2018) used a modified document retrieval component of

[1]The results on the test set and current rank are omitted to protect double blind review.

DrQA (Chen et al., 2017) to select the top most similar sentences w.r.t the claim, using bigram TF-IDF with binning.

- **Textual Entailment**: For the entailment task, training is done using labeled claims paired with evidence (labels are Supported, Refuted, NotEnoughInfo). Thorne et al.(2018) used the decomposable attention model (Parikh et al., 2016) for this task. For the case where multiple sentences are required as evidence, the strings were concatenated.

Our system implements changes in all three modules (Section 2), which leads to significant improvements on the dev set. Our document retrieval approach covers 94.4% of the claims requiring evidence, compared to 55.30% in the baseline. Further, our evidence recall is improved by 33 points over the baseline. For entailment, our model improves the baseline by 7.5 points on dev set. Overall, our end-to-end system shows an improvement of 19.56 in FEVER score compared to the baseline (50.83 vs. 31.27) (Section 3). Together with the results we discuss some lessons learned based on our error analysis and release our code [2].

2 Methods

2.1 Document Retrieval

Document Retrieval is a crucial step when building an end-to-end system for fact extraction and verification. Missing a relevant document could lead to missed evidence, while non-relevant documents would add noise for the subsequent tasks of sentence selection and textual entailment. We propose a multi-step approach for retrieving documents relevant to the claims.

- **Google Custom Search API**: Wang et al. looked at retrieving relevant documents for fact-checking articles, looking at generating candidates via search. Inspired by this, we first use the Custom Search Api of Google to retrieve documents having information about the claim. We add the token `wikipedia` to the claim and issue a query and collect the top 2 results.

- **Named Entity Recognition**: Second, we use the AllenNLP (Gardner et al., 2017) pre-trained bidirectional language model (Peters et al., 2017) for named entity recognition [3]. After finding the named entities in the claim, we use Wikipedia python API [4] to collect the top wikipedia document returned by the API for each named entity.

- **Dependency Parse**: Third, to increase the chance of detecting relevant entities in the claim, we find the first lower case verb phrase (VP) in the dependency parse tree and query the Wikipedia API with all the tokens before the VP. The reason for emphasizing lower case verb phrase is to avoid missing entities in claims such as "Finding Dory was directed by X", where the relevant entity is "Finding Dory". To deal with entity ambiguity, we also add the token `film` in our query where the claim contains keywords such as `film`, `stars`, `premiered` and `directed by`. For example in "Marnie was directed by Whoopi Goldberg.", Marnie can refer to both wikipedia pages `Marnie (film)` and `Marnie`. Our point of interest here is `Marnie (film)`.

- **Combined**: We use the union of the documents returned from by the three approaches as the final set of relevant documents to be used by the Sentence Selection module.

Method	Avg k	Coverage
Google API	2	79.5%
NER	2	77.1%
Dependency Parse	1	80.0%
Combined	3	**94.4%**
(Thorne et al., 2018)	5	55.3%

Table 2: Coverage of claims that can be fully supported or refuted by the retrieved documents (dev set)

Table 2 shows the percentage of claims that can be fully supported or refuted by the retrieved documents before sentence selection. We see that our best approach (combined) achieved a high coverage 94.4% compared to the baseline (Thorne et al., 2018) of 55.3%.

2.2 Sentence Selection

For sentence selection, we used the modified document retrieval component of DrQA (Chen et al., 2017) to select sentences using bigram TF-IDF with binning as proposed by (Thorne et al., 2018).

[2] http://anonymous (link suppressed for anonymous reasons)

[3] http://demo.allennlp.org/named-entity-recognition

[4] https://pypi.org/project/wikipedia/

We extract the top 5 most similar sentences from the k-most relevant documents using the TF-IDF vector similarity. Our evidence recall is 78.4 as compared to 45.05 in (Thorne et al., 2018), which demonstrates the importance of document retrieval in fact extraction and verification.

However, even though TF-IDF proves to be a strong baseline for sentence selection we noticed on the dev set that using all 5 evidences together introduced additional noise to the entailment model. To solve this, we further filtered top 3 evidences from the selected 5 evidences using distributed semantic representations. Peters et al. (2018) show how deep contextualized word representations model both complex characteristics of word use (e.g., syntax and semantics), and usage across various linguistic contexts. Thus, we used the ELMo embeddings to convert the claim and evidence to vectors. We then calculated cosine similarity between claim and evidence vectors and extracted the top 3 sentences based on the score. Because there was no penalty involved for poor evidence precision, we returned all five selected sentences as our predicted evidence but used only top three sentences for the entailment model.

2.3 Textual Entailment

The final stage of our pipeline is recognizing textual entailment. Unlike Thorne et al. (2018), we did not concatenate evidences, but trained our model for each claim-evidence pair. For recognizing textual entailment we used the model introduced by Conneau et al. (2017) in their work on supervised learning of universal sentence representations.

The architecture is presented in Figure 1. We use bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with max-pooling to encode the claim and the evidence. The text encoder provides dense feature representation of an input claim or evidence. Formally, for a sequence of T words $w_{t=1,..,T}$, the BiLSTM layer generates a sequence of h_t vectors, where h_t is the concatenation of a forward and a backward LSTM output. The hidden vectors h_t are then converted into a single vector using max-pooling, which chooses the maximum value over each dimension of the hidden units. Overall, the text encoder can be treated as an operator Text $\rightarrow R^d$ that provides d dimensional encoding for a given text.

Out of vocabulary issues in pre-trained word

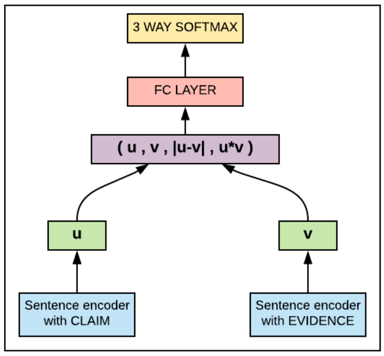

Figure 1: The architecture for recognizing textual entailment (adapted from (Conneau et al., 2017))

embeddings are a major bottleneck for sentence representations. To solve this we use fastText embeddings (Bojanowski et al., 2017) which relies on subword information. Also, these embeddings were trained on Wikipedia corpus make them an ideal choice for this task.

As shown in Figure 1, the shared sentence encoder outputs a representation for the claim u and the evidence v. Once the sentence vectors are generated, the following three methods are applied to extract relations between the claim and the evidence (i) concatenation of the two representations (u, v); (ii) element-wise product u*v and (iii) absolute element-wise difference $|u - v|$. The resulting vector, which captures information from both the claim and the evidence, is fed into a 3-class classifier consisting of fully connected layers culminating in a softmax layer.

For the final class label, we experimented first by taking the majority prediction of the three (claim, evidence) pairs as our entailment label but this led to lower accuracy on the dev set. So our final predictions are based on the following rule: if there is a SUPPORTS relation and the other two are NOT ENOUGH INFO the label is SUPPORTS; if there is a REFUTES relation and the other two are NOT ENOUGH INFO the label is REFUTES; otherwise we take the label that has the majority count. The accuracy of our entailment model is 58.77 on the dev set, which is 7.5 points higher than the baseline.

Implementation Details. The batch size is kept as 64. The model is trained for 15 epochs using Adam optimizer with a learning rate of 0.001. The size of the LSTM hidden units is set to 512 and for the classifier, we use a MLP with 1 hidden-layer of 512 hidden units. The embedding dimension of the words is set to 300.

3 E2E Results and Error Analysis

Table 3 shows the overall FEVER score obtained by our pipeline on the dev set of the Shared task (test set results withheld to preserve double blind review).

Data	Pipeline	FEVER
DEV	(Thorne et al., 2018)	31.27
	Ours	**50.83**

Table 3: FEVER scores on shared task dev set

On closer investigation we find that neither TF-IDF nor sentence embedding based approaches are perfect when it comes to sentence selection, although TF-IDF works better.

Fox 2000 Pictures released the film Soul Food	0.29
Soul Food is a 1997 American comedy-drama film produced by Kenneth "Babyface" Edmonds ,Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures	

Table 4: Cosine similarity between claim and supporting evidence

Table 4 goes on to prove that we cannot rely on models that entirely depend on semantics. In spite of the two sentences being similar, the cosine similarity between them is poor mostly because the evidence contains a lot of extra information which might not be relevant to the claim and difficult for the model to understand.

At seventeen or eighteen years of age, he joined Plato's Academy in Athens and remained there until the age of thirty-seven (c. 347 BC)
Shortly after Plato died , Aristotle left Athens and at the request of Philip II of Macedon ,tutored Alexander the Great beginning in 343 BC

Table 5: The top evidence is selected by Annotators and the bottom evidence by our pipeline

We also found instances where the predicted evidence is correct but it does not match the gold evidence. For the claim "Aristotle spent time in Athens", both evidences given in Table 5 support it, but still our system gets penalized on not being able to match the gold evidence.

We found quite a few annotations to be incorrect and hence the FEVER scores are lower than expected. Table 6 show two instances where the gold labels for the claims was NOT ENOUGH INFO, while in fact they should have been SUPPORTS and REFUTES, respectively.

Claim: Natural Born Killers was directed by Oliver Stone **Evidence**: Natural Born Killers is a 1994 American satirical crime film directed by Oliver Stone and starring Woody Harrelson , Juliette Lewis , Robert Downey Jr. , Tom Sizemore , and Tommy Lee Jones .
Claim:Anne Rice was born in New Jersey **Evidence**: Born in New Orleans, Rice spent much of her early life there before moving to Texas, and later to San Francisco

Table 6: Wrong gold label (NOT ENOUGH INFO)

Table 7 reflects the fact that NOT ENOUGH INFO is often hard to predict and that is where our model needs to improve more.

	S	N	R
S	**4635**	1345	686
N	2211	**3269**	1186
R	1348	1470	**3848**

Table 7: Confusion matrix of entailment predictions

The lines between SUPPORTS and NOT ENOUGH INFO is often very blurred as shown in Table 7. Our models needs better understanding of semantics to be able to identify these. Table 8 shows one such example where the `gospel` keyword becomes the discriminative factor.

Claim: Happiness in Slavery is a gospel song by Nine Inch Nails
Evidence: Happiness in Slavery,is a song by American industrial rock band Nine Inch Nails from their debut extended play (EP), Broken(1992)

Table 8: Example where our model predicts SUPPORTS for a claim labeled as NOT ENOUGH INFO

4 Conclusion

The FEVER shared task is challenging primarily because the annotation requires lot of dedication and manual effort. We presented an end-to-end pipeline to automate the human effort and show empirically that our model outperforms the baseline by a large margin. We also provided a thorough error analysis which highlighted some of the shortcomings of our models and potentially of the gold annotations.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. pages 1870–1879. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Oyvind Tafjord Mark Neumann, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. In *Neural computation, 791*, pages 1735–1780.

Ankur Parikh, Dipanjan Das Oscar Tackstr om, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. pages 2249–2255. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. pages 809–819. Proceedings of NAACL-HLT 2018.

Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. pages 525–533. WWW '18 Companion Proceedings of the The Web Conference 2018.