Team UMBC-FEVER : Claim verification using Semantic Lexical Resources

Ankur Padia, Francis Ferraro and Tim Finin
Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 20150 USA
{pankur1,ferraro,finin}@umbc.edu

Abstract

We describe our system used in the 2018 FEVER shared task. The system employed a frame-based information retrieval approach to select Wikipedia sentences providing evidence and used a two-layer multilayer perceptron to classify a claim as correct or not. Our submission achieved a score of 0.3966 on the Evidence F1 metric with accuracy of 44.79%, and FEVER score of 0.2628 F1 points.

1 Introduction

We describe our system and its use in the FEVER shared task (Thorne et al., 2018). We focused on two parts of the problem: (i) *information retrieval* and (ii) *classification*. For the first we opted for a linguistically-inspired approach: we automatically annotated claim sentences and Wikipedia page sentences with syntactic features and semantic frames from FrameNet (Baker et al., 1998a) and used the result to retrieve sentences relevant to the claims that provide evidence of their veracity. For classification, we used a simple two-layer perceptron and experimented with several configurations to determine the optimal settings.

Though the overall classification of our best version was lower than the best approach from Thorne et al. (2018), which used a more sophisticated classification approach, we scored 10^{th} out of 24 for the information retrieval task (measured by F_1). The improvement in our system worked well on the IR task, obtaining a relative improvement of 131% on retrieving evidence over the baseline F1 measure Thorne et al. (2018).

2 Approach

The FEVER task requires systems to assess a sentence making one of more factual claims (e.g., "Rocky Mountain High is an Australian song") as true or false by finding sentences in Wikipedia that provide evidence to support or refute the claim(s). This naturally leads to two sub-tasks, an *information retrieval* task that returns a set of Wikipedia sentences that are relevant to the assessment and a *classification* task that analyzes the evidence and labels the claim as Supported, Refuted or NotEnoughInfo. Figure 1 shows the overall flow of our system, which uses semantic frames to analyze and match a claim sentence to potential evidence sentences and a multilayer perceptron for claim classification.

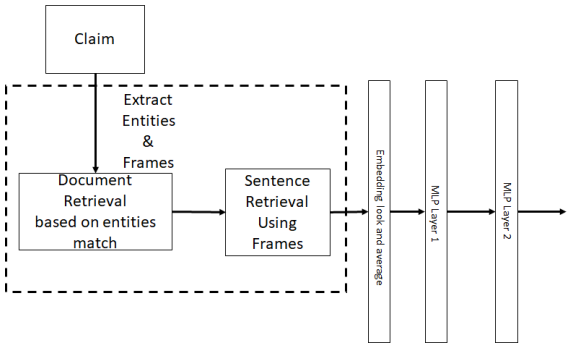

Figure 1: Our system used a semantic frame approach to support both retrieval of claim evidence and classification of claims.

2.1 Finding Relevant Evidence Sentences

Our approach used semantic frames from FrameNet (Baker et al., 1998b) as part of the analysis in matching a claim with sentences that might provide evidence for its veracity. A frame is a semantic schema that describes a situation, event or relation and its participants. The FrameNet collection has more than 1,200 frames and 13,000 lexical units which are lemmas that evoke or trigger a frame; see Fig. 2 for an example of this schema. Complex concepts and situations can be described by multiple frames. As an example, the sentence 'John bought a new

Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 161–165
Brussels, Belgium, November 1, 2018.

Who	Classifier	Training type	Classification		Predicting evidence		
			FEVER Score	ACC	Precision	Recall	F1
$UMBC_1$	MLP	**NFC**	0.2572	0.4398	0.4868	0.3346	0.3966
$UMBC_2$	MLP	**NFUC**	0.2628	0.4479	**0.4868**	**0.3346**	**0.3966**
$UMBC_3$	MLP	**NFIC**	0.2599	0.4069	0.4868	0.3346	0.3966
$Baseline_1$	MLP	(Thorne et al., 2018, Tab. 4)	0.1942	40.64	–	–	–
$Baseline_2$	DA	CodaLab results	**0.3127**	**0.5137**	–	–	0.1718

Table 1: Performance on development dataset of the system on different settings. We achieve comparable classification performance with simple classifier model thanks to better evidence retrieval.

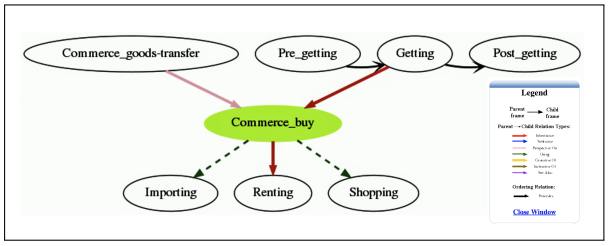

Figure 2: The FrameNet Commerce_buy frame and its immediate neighbors

bike' can trigger two frames: 'Claim_ownership'[1] and 'Commerce_buy'.[2]

Our annotation processing differed slightly for claims and potential evidence. We processed the claims in the dataset with the annotation pipeline described in Ferraro et al. (2014). Each claim was annotated using a named entity recognizer, dependency parser and POS tagger from CoreNLP (Manning et al., 2014) and also by a frame annotator (Das et al., 2010). For the evidence sentences, we used the pre-existing semantically annotated version of Wikipedia (Ferraro et al., 2014) that contained the same types of annotations for all of Wikipedia pages from a 2016 Wikipedia dump, serialized as Thrift objects using the *concrete* schema (HLTCOE, 2018).

Depending on the dataset, we performed document and sentence retrieval. We did only sentence retrieval for the training data and for development and testing data we did document and sentence retrieval. Our motivation was to understand the effect of frame-based retrieval, assuming the named entity recognizer correctly identified the entity.

For the training dataset, we used the dataset's document titles to retrieve Wikipedia documents directly and choose its sentences that triggered some frame as candidate evidence sentences. We applied exact frame matching, in which a sentence is predicted as evidence if it triggers a frame that is also triggered by the claim. All sentences that had an exact match to one of the claim's frames were added to the final evidence set.

The extraction of the documents and the sentences was different in the case of the development and testing datasets as no gold standard document identifiers were available. We used a two-layer multilayer perceptron to label the claim given the evidence as either SUPPORTS, REFUTES or NotEnoughInfo.

One complication was that the evidence was extracted from a different Wikipedia dump (2017) than our frame-annotated Wikipedia corpus (2016). While the page title's aligned well between the two Wikipedia dumps, their sentences exhibited more variations. This is the result of Wikipedia editors making changes to the page, including rearrangements, updates, adding material or stylistic modifications. In order to find the correct sentence index in the page, we used the Hungarian algorithm (Kuhn, 1955) to find the matching sentences. We cast this problem as a dissimilarity minimization problem, where the dissimilarity between a pair of sentences was 1 minus a Jacard similarity metric over the set of sentence tokens.

2.2 Classifying Claims Given the Evidence

To produce features, we converted each claim word to a 400 dimension embedding (Mikolov et al., 2013) representation and took the average over the length of the claim, using a zero vector for out-of-vocabulary words. We trained a two-layer MLP to label the claim using stochastic gradient descent with L2 and dropout to avoid overfitting. We chose the final parameter values for the claim classifier that gave best result on development dataset, which are shown in Table 2.

[1]The *Claimant* asserts rights or privileges, typically ownership, over some *Property*. The *Claimant* may be acting on the behalf of a *Beneficiary*.

[2]The *Buyer* wants the *Goods* and offers *Money* to a *Seller* in exchange for them.

Parameter	Value
learning rate	0.01
number of layers	2
optimizer	SGD
hidden layer size	50
L2 regularize	1e-06
epoch	2
batch_size	64
dropout	0.5

Table 2: MLP classifier parameter values

3 Ablation Study

We explored our approach by evaluating performance with settings corresponding to three different information retrieval strategies.

- **NFC**: NER document retrieval + Frame sentence retrieval + Classification
- **NFUC**: NER document retrieval + Frame sentence retrieval + (Union) introduction section of the Wikipedia page (Thorne et al., 2018) + Classification
- **NFIC**: NER document retrieval + Frame sentence retrieval + (intersection) introduction section of Wikipedia page + Classification

3.1 Results

Table 1 shows confusion matrices of our system when trained with the three different settings. The performance to predict the score is the same as we are retrieving the frame based sentences from the documents and adding FEVER processed Wikipedia sentences on the fly at the training time. The addition of FEVER-processed Wikipedia sentences slightly increases the performance of the system.

Since the frame annotator is not perfect, it sometimes fails to trigger appropriate frames. This means that while the vast majority of claims could be matched with potential evidence, there are claims that cannot be matched with evidence. This was neither uncommon nor rare: in the development set, 21.43% of the claims could not be matched with evidence sentences (the testing and training datasets had miss rates of 17.43% and 25.78%, respectively).

As evident from the classification performance, additional data improves performance, with NFUC performing better than other two settings. Compared to the results in the test dataset, we scored nearly twice as well as the baseline in terms of information retrieval with simple frame matching. This is evidence for the effectiveness of using semantic frames in determining the credibility of the claim, despite the recall issues discussed above.

		Predicted		
		Support	**Refute**	**Neither**
Actual	**Support**	4646	171	1849
	Refute	3050	1198	1618
	Neither	4123	391	2152

(a) Dev confusion matrix for frame-based sentence retrieval only (NFC).

		Predicted		
		Support	**Refute**	**Neither**
Actual	**Support**	4499	173	1994
	Refute	2777	2125	1764
	Neither	3968	365	2333

(b) Dev confusion matrix for the union of frame-based and introduction-based sentence retrieval (NFUC).

		Predicted		
		Support	**Refute**	**Neither**
Actual	**Support**	4370	87	2209
	Refute	3122	1474	2070
	Neither	4159	214	2293

(c) Dev confusion matrix for the intersection of frame-based and introduction-based sentence retrieval (NFIC).

Table 3: Classification-without-provenance accuracy confusion matrices on the development dataset for the three classes under.

3.2 Discussion

The three settings had similar performance measures because the set of sentences found by our system was a superset of those found by the human assessors. Our frame-based retrieval found 516,670 evidence sentences when matching frames across the entire document mentioning entities and not just the introduction section. The set found by assessors included 34,797 evidence sentences, all of which were included the frame-based retrieval set.

Fig. 3 shows a correct and incorrect example. A manual examination revealed that the predicting evidence was correct nearly every time when an appropriate frame was in the document. When a frame is in the claim and not in the document,

***Claim*: Last Man Standing does not star Tim Allen**
Predicted evidence (Correct):

1. Timothy Allen Dick (born June 13, 1953), known professionally as Tim Allen, is an American actor, comedian and author
2. He is known for his role as Tim "The Toolman" Taylor in the ABC television show Home Improvement (1991) as well as for his starring roles in several films, including the role of Buzz Lightyear in the Toy Story franchise
3. From 2011 to 2017, he starred as Mike Baxter in the TV series Last Man Standing

***Predicted Label*: REFUTES (due to evidence (3))**
***Actual Label*: REFUTES**

(a) Relevant evidence is correctly retrieved and is classified correctly as refuting the claim.

***Claim*: Rocky Mountain High is an Australian song**
Predicted evidence (Correct):

1. "Rocky Mountain High" is a folk rock song written by John Denver and Mike Taylor about Colorado, and is one of the two official state songs of Colorado
2. The song also made #3 on the Easy Listening chart, and was played by some country music stations
3. Denver told concert audiences in the mid-1970s that the song took him an unusually long nine months to write
4. Members of the Western Writers of America chose it as one of the Top 100 Western songs of all time

***Predicted Label*: SUPPORTS**
***Actual Label*: REFUTES**

(b) Relevant evidence is correctly retrieved, but was misclassified by the classifier as supporting the claim.

Figure 3: Error analysis examples of predicting evidence and classification. As evident from the examples, the frame based retrieval extracts high quality evidence sentences when available in the document. However, the performance of the system is reduced depending on the classifier predictions, and perfection of the automatic frame annotator.

the retrieval component gives empty results and, depending on the gold standard, the performance suffers. The mismatch happens due to differences with the Wikipedia version dumps. The FEVER dataset used a 2017 dump and we used one from 2016. A second error source was annotation/misclassification by our frame annotation system. However, whenever there is a match, the quality of the evidence is high, as shown by the first and second example claims. Table 3 shows the confusion matrices for the three classes (Support, Refute, Not enough information) for each of the three settings.

4 Discussion and Conclusion

Our submission was an initial attempt to explore the idea of using semantic frames to match claims with sentences providing evidence that might support or refute them. The approach has the advantage of being able to exploit relations between frames such as entailments, temporal ordering, causality and generalization that can capture common sense knowledge. While the classification scores were lower than we hoped, the evidence retrieval scores represent impressive and promising improvements.

We plan to continue developing the approach and add it as a component of a larger system for cleaning noisy knowledge graphs (Padia, 2017; Padia et al., 2018).

We expect that the performance measures will improve when the datasets are all extracted from the identical Wikipedia versions. Possible enhancements include using the Kelvin (Finin et al., 2015) information extraction system to add entity coreference and better entity linking to a knowledge graph of background knowledge, such as Freebase, DBpedia or Wikidata. This will sup-

port linking nominal and pronominal mentions to a canonical named mention and provide access to more common aliases for entities. Such features have been shown to improve entity-based information retrieval (Van Durme et al., 2017).

We also hope to exploit Kelvin's ability to reason about entities and relations. Its knowledge graph knows, for example, that while one can only be born in single geo-political location, such places are organized in a *part-of* hierarchy. An event that happens in one a place can be said to also take place at its enclosing locations. The system's background knowledge includes that *Honolulu* is part of *Hawaii* which in turn is part of the *United States*. Moreover, it knows that if you were born in a country, it is very likely that you are a citizen of that country. This will allow it to recognize "Obama was born in Honolulu" as evidence that supports the claim that "Obama is a citizen of the U.S.".

Acknowledgments

This research was partially supported by gifts from the IBM AI Horizons Network and Northrop Grumman and by support from the U.S. National Science Foundation for UMBC's high performance computing environment.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998a. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 86–90. Association for Computational Linguistics.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998b. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A Smith. 2010. Probabilistic frame-semantic parsing. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 948–956. Association for Computational Linguistics.

Francis Ferraro, Max Thomas, Matthew R Gormley, Travis Wolfe, Craig Harman, and Benjamin Van Durme. 2014. Concretely annotated corpora. In *AKBC Workshop at NIPS*.

Tim Finin, Dawn Lawrie, Paul McNamee, James Mayfield, Douglas Oard, Nanyun Peng, Ning Gao, Yiu-Chang Lin, Josh MacLin, and Tim Dowd. 2015. HLTCOE participation in TAC KBP 2015: Cold start and TEDL. In *Text Analytics Conference (TAC)*.

HLTCOE. 2018. Concrete. http://hltcoe.github.io-/concrete/.

Harold W. Kuhn. 1955. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, page 8397.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

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*, pages 3111–3119.

Ankur Padia. 2017. Cleaning Noisy Knowledge Graphs. In *Proceedings of the Doctoral Consortium at the 16th International Semantic Web Conference*, volume 1962. CEUR Workshop Proceedings.

Ankur Padia, Frank Ferraro, and Tim Finin. 2018. KG-Cleaner: Identifying and correcting errors produced by information extraction systems. *arXiv preprint arXiv:1808.04816*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. *CoRR*, abs/1803.05355.

Benjamin Van Durme, Tom Lippincott, Kevin Duh, , Deana Burchfield, Adam Poliak, Cash Costello, Tim Finin, Scott Miller, James Mayfield, Philipp Koehn, Craig Harmon, Dawn Lawrie, Chandler May, Max Thomas, Annabelle Carrell, and Julianne Chaloux. 2017. CADET: Computer Assisted Discovery Extraction and Translation. In *8th International Joint Conference on Natural Language Processing (System Demonstrations)*, pages 5–8.