CSE 635

Natural Language Processing and Text Mining

Milestone 2 - Interim Report

Fact Extraction and Automated Claim Verification

Team: tmp

Members:

Anas, Anirudh, Snigdha

Data Set

The dataset built for this challenge consists of claims that have already been labeled by human annotators. The dataset is split into ~*79%* training data, ~*5%* development data, ~*5%* test data and ~*11%* held out data. The table below, which was taken from the FEVER challenge paper [1], gives the exact figures. Each sample in the dataset contains a claim with the following fields:

Split	Supported	Refuted	NEI
Training	80035	29775	35639
Dev	3333	3333	3333
Test	3333	3333	3333
Reserved	6666	6666	6666

Training Data	
Id	150448
Verifiable	VERIFIABLE
Label	SUPPORTS
Claim	The 1991 NBA Finals had the Lakers in it.
Evidence	[[36004, 43545, "1991_NBA_Finals", 8]]
	[[36004, 43546, "1991_NBA_Finals", 14]]
	[[36004, 43547, "1991_NBA_Finals", 17]]
Unlabelled Data	
Id	194462
Claim	Tilda Swinton is a vegan.

Wikipedia Evidence	
id	1958–59_NBA_season
text	The 1958 -- 59 NBA Season was the 13th season of the National Basketball Association . The season ended with the Boston Celtics winning the NBA Championship -LRB- the first of what would be 8 straight -RRB- , beating the Minneapolis Lakers 4 games to 0 in the NBA Finals.
lines	The 1958 -- 59 NBA Season was the 13th season of the National Basketball Association. National Basketball Association National Basketball Association Basketball Basketball
	The season ended with the Boston Celtics winning the NBA Championship -LRB- the first of what would be 8 straight -RRB- , beating the Minneapolis Lakers 4 games to 0 in the NBA Finals. Boston Celtics Boston Celtics Minneapolis Lakers Minneapolis Lakers NBA Finals 1959 NBA Finals

Figure 1 FEVER Data Set

Architecture

Document Retrieval: It is defined as the matching of some stated user query against a set of free-text records[2]. For the baseline implementation, a PyTorch implementation of DrQA [3] is used. We create a simple inverted index lookup and perform term-vector model scoring. We then compare the claim and evidence as TF-IDF weighted bag-of-words vectors.

Sentence Selection: It is the task of identifying sentences that contain the answer to a given question or claim [4].

In order to verify a claim, the project aims at selecting specific sentences from the retrieved documents to support or refute the claim. We extract the top 5-most similar sentences from the 5-most relevant documents using TF-IDF similarity. For TF-IDF, PyTorch implementation of DrQA is used. First, we rank sentences based on TF-IDF similarity to the claim. Then sort most similar sentences first and tune a cut-off using validation similarity on the development set. Finally, we get top results using cosine similarity based on the above scores.

Recognizing Textual Entailment: RTE is a generic task of capturing major semantic inference needs in Natural Language Processing applications. This means that given two fragments of text, recognize whether the meaning of one fragment of text is entailed (can be inferred) from another text.[5]

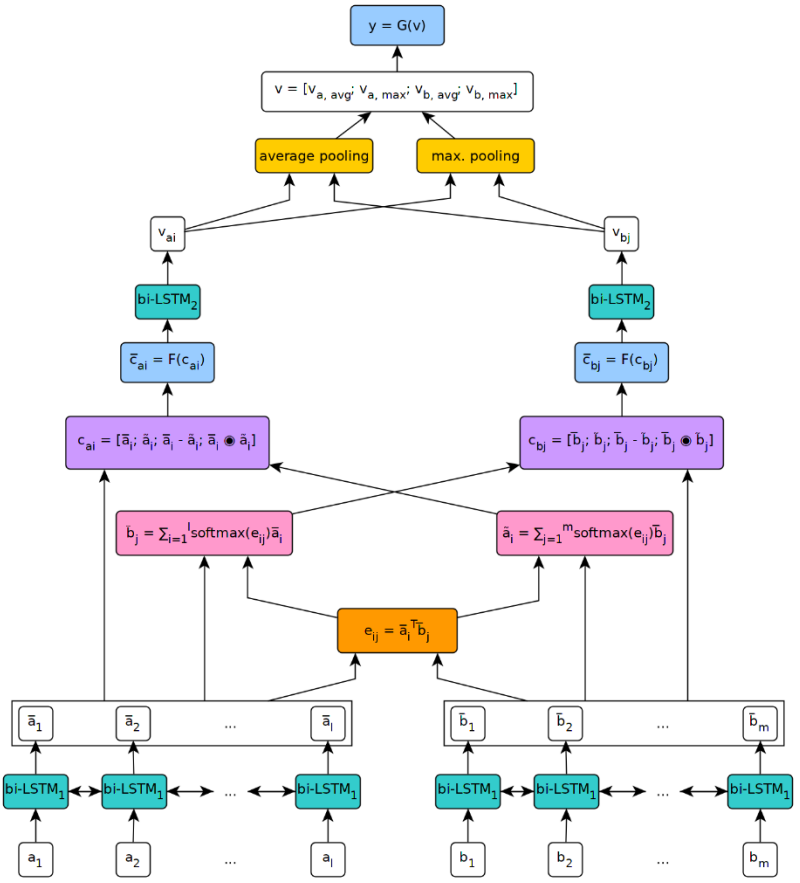

Figure 2 Block diagram for ESIM

In order to perform RTE, ESIM[6] (Enhanced Sequential Inference Model) was used, which primarily uses two bi-LSTM layers to encode the context, which is provided as the claim and the evidence sentences, and evaluates the relationship between the encoded input through a Linear layer. The input is first represented as a correspondence to vectors using the GloVe embedding.

The figure on the right depicts the input as it flows through the first bi-LSTM layer, undergoing a series of transformations to yield an augmented representation of the intermediate output. The output then is fed to a second bi-LSTM layer, the output of which is projected onto a single vector. The path culminates with that aggregated vector being fed to a linear layer, which in turn produces a probability for each of the three labels (*'Supports'*, *'Refutes'*, and *'Not Enough Info'*).

Evaluation

The measurement of the quality of predictions takes into account how well each of the three major components in the system performs. For each claim, there is a set of evidence groups. Each evidence group contains ground truth evidence sentences. The scoring mechanism discussed below considers each evidence group separately until either a matching group is found, or the prediction fails to provide a complete group of evidence sentences in any of the ground truth sets.

For document retrieval and sentence selection, the results are compared with the documents and sentences selected by the ground truth and precision, recall and F1 score is then calculated for "Supported" and "Refuted" claims and are ignored in the case of "Not enough info". In RTE for the *'Supports'* or *'Refutes'* labelled samples, the gold evidence was used directly in the training cycle by concatenating all the sentences in the longest evidence set, and feeding them as one of the two inputs to the first bi-LSTM layer. For *'Not Enough Info'*, labeled samples a preprocessing step was performed where, for each such sample, the best scoring 5 sentences from the closest 5 documents in terms of similarity to the claim were fetched using the methods described above. These 5 sentences were then used as negative evidence to indicate to the ESIM network what resembles a case where the provided evidence is not sufficient to make a clear decision. RTE is measured through the accuracy of the label classifications for a given set of claims.

Additionally, the FEVER score is used where we submit the final results from the application. The fever score is calculated on the reserved set where the claims with correct labels and evidence versus the total number of claims.

Results

Document retrieval and sentence selection were carried out using a simple TF-IDF and cosine similarity method. However, for ESIM network was trained on the provided training and development datasets, which consist of a series of labeled claims along with their gold evidence sets. While training, the accuracy of the RTE system in terms of when it makes a correct label classification (regardless of the precision of the evidence set) was tracked, and the results can be shown in the graph below.

In addition to the training results, the predictions for the provided test set were submitted to the official ranking system made available by the FEVER competition organizers. The results that were obtained were a slight improvement from the FEVER score of the baseline system, which was around *27%,* to *33%* using the ESIM network.

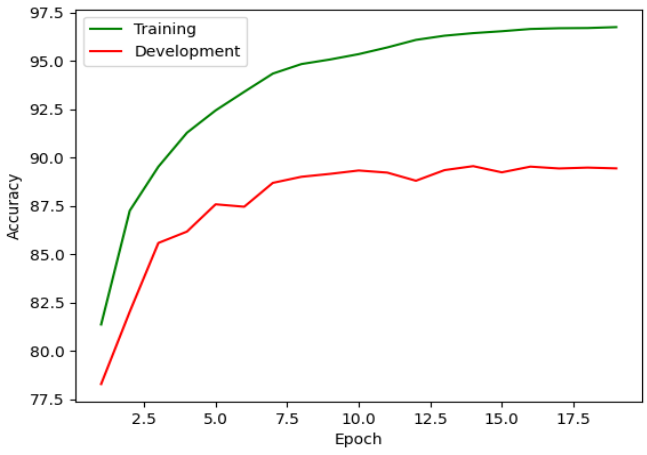

Figure 3 Accuracy for label classification (RTE system), trained for 20 epochs

Metric	Baseline Dataset Score [3]	Test Dataset Score (our system)
FEVER Score	27.45%	**33.5%**
Label Accuracy	48.84%	34.33%
Evidence Precision	11.28%	0.35%
Evidence Recall	47.87%	1.43%
Evidence F1	18.26%	0.57%

Metric	Baseline Dataset Score [3]	Dev Dataset Score (our system)
Label Accuracy	48.84%	**89%**
Evidence Precision	11.28%	**24.85%**
Evidence Recall	47.87%	21.33%
Evidence F1	18.26%	**22.96%**

Table 1 Test and Dev Dataset scores for our system

Proposed Plan

Document Retrieval

The current system makes use of TF-IDF between the document and the claim.

In the proposed system, NER and noun phrase extraction will be used to find key topics in the claims and documents helping improved results. The matched documents would be ranked using TF-IDF. Initial version of NER will be implemented using Spacy and can be further improved using a pre-trained model such as BERT. TF-IDF will be calculated using the DrQA model as it is being used currently. Additionally, topic modelling can be used on each document to provide better disambiguation. Topic modelling can be performed using Mallet. Further improvements can be performed by using disambiguation given in the article ID to be used as meta information.

Sentence Selection

The current system uses TF-IDF and cosine similarity to select the sentences. In the proposed system, the position of sentence, length of document, keywords would be passed to a multilayer perceptron in addition to the claim and the sentence and returns a score. Top k sentences are selected and passed to the next layer in the pipeline. Additionally, we will also experiment with Cross-attention Convolutional Layer to see if the system can handle simple word co-occurrence criteria and see if results can be enhanced [7].

RTE

Going forward with the ESIM model, the network will be kept the same, with the possibility of augmenting some of the layers with self-attention. Additionally, neural aggregation will be explored, where instead of concatenating all the evidence sentences together into one sentence, they could be evaluated separately, and have a linear layer draw a decision on the claim based on the logits provided by the neural aggregation layer. This strategy might prove more useful, as it maintains the integrity of each sentence, allowing for the context of each sentence to be factored into the decision without having it impacted directly by the other sentences, as is the case with the implementation discussed here.

References

[1] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The Fact Extraction and VERification (FEVER) Shared Task. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

[2] https://en.wikipedia.org/wiki/Document_retrieval

[3] https://github.com/facebookresearch/DrQA

[4] https://www.researchgate.net/publication/269117118_Deep_Learning_for_Answer_Sentence_Selection

[5] https://aclweb.org/aclwiki/Recognizing_Textual_Entailment

[6] *Chen Q, Zhu X, Ling Z, Wei S, Jiang H., Inkpen D. 2017. Enhanced LSTM for Natural Language Inference. arXiv:1609.06038v3 [cs.CL]. https://arxiv.org/abs/1609.06038*

[7] [PDF] Cross Attention for Selection-based Question Answering