Automatic Fact Verification

Site Huang 908282
Yizhou Wang 669026

1. Introduction

The fact verification assesses the truthfulness of a claim and this time, the system building depends upon the provided wiki page text to classify each of the statement. Any related information is retrieved to be compared with the claim and there are three categories settled down in the label feature, support, refute and not enough info.

2. Datasets

Wiki-text is used as the information referred to find the best match document to categorize the label. Training the model is based on the Training data with approximately 150K instances. Devset data further develops the model between the claim and label. Lastly, a developed model is applied to predict the claim in the test data.

3. Preprocess

The presentation format is obvious but there are still some preprocesses needed to utilize. For example, the claim and label are re-organized in order to meet the analyzer's expectation.

Remove the stopwords, punctuation, and verb words but carefully keep the negate such as not, n't; Lowering case for each word.

Original claim: Roman Atwood is a content creator.

Processed claim: roman atwood content creator

4 Fact Verification System

In this section, a fact verification system is introduced, which comprises of doc retrieval, sentence retrieval, and label prediction.

4.1 Doc Retrieval

Firstly, the query is passed into a Doc retrieval system. The Doc retrieval system is designed to retrieve the top-k documents that best match the query (claim) [1]. A Doc retrieval system comprises of the following two components.

Candidate documents search:

The main objective of searching candidate documents is to ensure the potential correct documents returned in a ranking list.

- Pylucene tool is used for text indexing and searching relevant documents. Indexing files are constructed by storing the content of the document. Several sentences with the same title are considered as a document in the Wiki corpus.
- Moreover, a custom analyzer is designed which has the capabilities of tokenizing sentence, lowering words, and filtering stopwords. A custom stopwords set including negative words (eg. 'NOT' words) are defined and used in the analyzer.
- In the meantime, BM25 similarity is used for information retrieval. TF-IDF with BM25 similarity is used for scoring and ranking the searching documents.

Keywords extraction:

The engine is integrated with a keywords extraction module in order to analyze and extract information from the query.

- The query sentence is parsed by using the spaCy tool. Every word is tagged by using part-of-speech tagging method. The words tagged with 'PROPN' (proper noun) are selected. The reason for choosing 'PROPN' tagged words is that these words basically can form together as a chunk, which contains most of the information of a sentence.
- For example, a claim "*Rachel Green was played by Janet Jackson.*" contains the information chunks "*Rachel Green*" and "*Janet Jackson*". These information chunks

are as keywords to help the search engine return more relevant documents.

4.2. Sentence Retrieval

The characteristics of sentence retrieval are implemented using Word2Vec and Cosine similarity. In addition, another method called Glove has been attempted and the detailed explanation is given to support for our choice.

Word2Vec utilizes different concepts comparing to the Cosine similarity as it tends to encode the word into the vector format provided by the word co-occurrence information. It is based on the concept of distributed presentations of words and phrases [1]. Thanks to the skip-gram and the CBOW models, any contexts or the specific word can be predicted given that the word in the middle or surrounding contexts are known. In this project, it is used to decide how similar does the claim sentence compared with the other one in the selected document because Word2Vec has the capacity to measure the relatedness between words. The vectors are obtained from Google news model and loaded into the Gensim module to calculate the sentence similarity. After removing the word that isn't in the pre-trained model and applying the preprocess, the score is becoming much more accurate for the decision.

Likewise, GloVe learns geometrical encodings of words from their co-occurrence information as well but it is a count-based model [2]. It takes consideration upon the whole picture rather than a single word window in Word2Vec.

The last technique that has been tried is the Cosine similarity. The first step is intuitive as the text is meaningless to the computer so that the text must be transformed into vectors [4]. Then the cosine similarity score is returned to indicate the angle between the sentences. The smaller the angle is, the more similar they are.

These three methods offer strong ability to select the most relevant evidence. They have outstanding performance without taking long-time training.

4.3. Label Prediction

The AllenNLP Textual Entailment (TE) model is used in this module. It takes a pair of sentences and predicts whether the facts in the first necessarily imply the facts in the second one [3]. As several candidates of evidence are selected in Sentence Retrieval module, each evidence is passed to the TE model to generate prediction along with the claim. The most likely label is selected eventually by analyzing label prediction results.

5. Enhancement

5.1 Doc Retrieval:

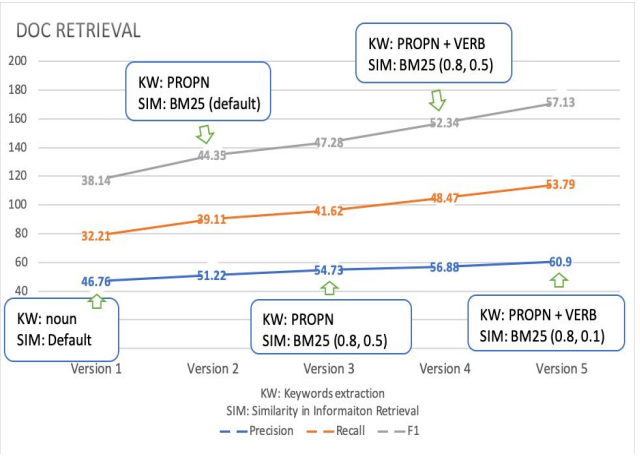

Figure 1: Document Retrieval

- In order to maximize the effectiveness of doc retrieval engine, two parameters of BM25 are set as *k=0.8* and *b=0.1*, respectively. The design decision of BM25 parameters is to reduce the penalty of frequent terms and the long length of the document as much as possible. Therefore, different parameters have been adopted in several experiments. And it shows that *k=0.8* and *b=0.1* have the best performance.
- Keywords extraction are tried in different ways. The first version tries to extract keywords by noun chunks, but it only

achieves 38.14% F1 scores. Version 2 uses 'PROPN' tagged words to find information chunks, which has achieved 6% improvement compared to version 1. Version 4 utilizes 'VERB' as an indicator to split information chunks, which shows 5% of improvement compared to version 3.

5.2 Sentence Retrieval:

Similarity Model:

Figure 2 compares the performance produced by three methods or related combinations mentioned before.

w&g combines Word2Vec and GloVe model
w&c combines Word2Vec and Cosine similarity

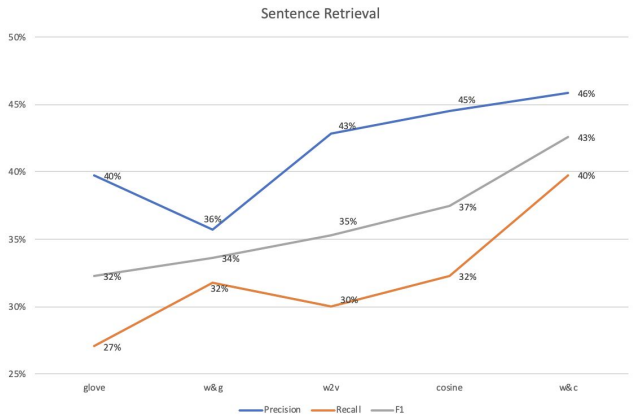

Figure 2: Sentence Retrieval

The GloVe has the poorest performance comparing to the other two. Although it is possible to bring some benefits when averaging all of the three methods, it brings a lot of workloads, especially it will take even longer to get the final prediction. In this way, GloVe is abandoned.

In the other way around, the chart statistically advises us to select w&c model because it gives the highest outcome for the sentence retrieval. In addition, the combination of both two models supplies diverse benefits to reduce the prediction loss effectively. Therefore, the combined method has the best performance.

Top-K Strategy:

The original strategy for sentence selection is made according to the maximum likelihood of the similarity score. However, it is not practical and could not chase the expectation. Therefore, a threshold is set in order to maximize the correctness. And according to distribution diagram shown below, the mean is quite close to 2 (1.86). Therefore, it is optimal to choose two sentences among the candidates of document as the threshold.

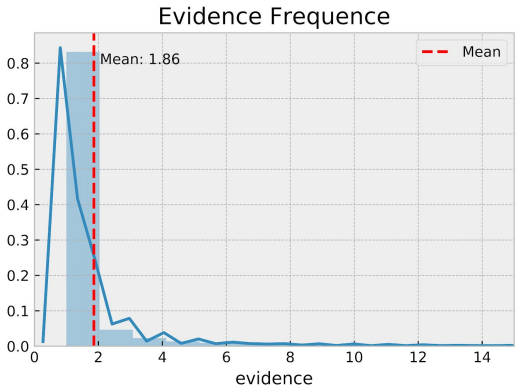

Figure 3: Evidence Frequency

Notably, the rise in the score is nearly 10% given that the same method has been applied and the only change is made to the top-K.

Best Evidence (TopK)	Sentence Selection (F1 score)
1	34.1%
2	43.1%

Table 1: Sentence Improvement

5.3 Label Prediction:

As the sentence retrieval engine return several sentences to the Text Entailment model, each sentence gets a prediction of labels according to the claim. As shown in Table 2. Version 1 uses the most common label as the final prediction, while version 2 calculates the average probability of labels and return the most likely label. Version 2 achieves 3% of improvement compared to version 1.

Versions	Label Accuracy
1	44.34 %
2	47.35 %

Table 2: Label Improvement

6. Error Analysis

In this section, the error analysis is discussed.

Claim: The State of Palestine claims Algeria.	
Methods	**Keywords**
Actual title	State_of_Palestine
Noun	The State; Palestine; Algeria
PROPN	Algeria; State_of_Palestine_claims_Algeria; Palestine; State;
PROPN + VERB	Algeria; State_of_Palestine

Table 3: Error Example

6.1 Doc retrieval:

Different extracting keywords methods are compared in Table 3, it shows noun chunks sometimes split the meaning of the sentence. 'PROPN' method often produces error keywords. Therefore, 'PROPN + VERB' method is proposed by using 'VERB' tagged word to split the keywords. The result shows it contains the correct keyword.

6.2 Sentence retrieval:

The most common errors made in the sentence retrieval is that the model could easily underestimate the similarity score when the entity in the claim does not appear in the annotated sentence or appear in the form of pronoun. For instance, the claim expresses that "Roman Atwood is a content creator" and the evidence for this claim is "He is best known for his vlogs...", "He also has another YouTube channel...". Neither of the sentences covers the entity's full name but use pronoun to present, in this way, the mistaken is easy to be made because the model cannot identify this pattern effectively.

6.3 Spelling errors:

There is an example that misspelled in the claim or in the evidence. "claim": "Homer Hickman is a writer of historical fiction books." responds to "Homer_Hickam" in the evidence. The system does not consider the scenario like this so it cannot find the document as expected.

6.4 Indirect document:

The claim may possess indirect connection with the other documents. For the claim "Lily James has been on TV.", it refers to the evidence not only "Lily_James", but also "Downton_Abbey" and "Just_William_(2010_TV_series)". The potential relationship is hard to be extracted by the system.

7. Codalab Results:

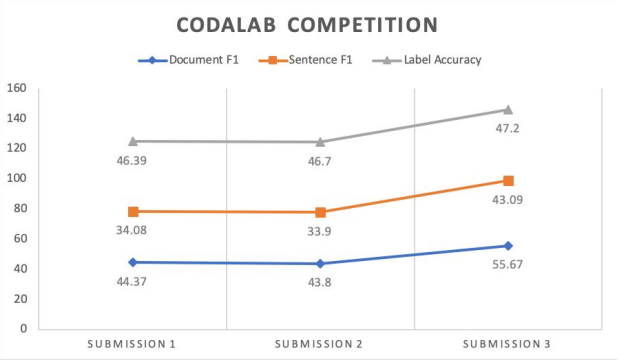

Figure 4: Competition Result

Codalab submissions are shown in Figure 4.

Submission 1: Version 3 of doc retrieval engine and w2v model for selecting sentences are used.
Submission 2: Version 3 of doc retrieval engine and w&g model for selecting sentences are used.
Submission 3: Version 5 of doc retrieval engine and w&c model for selecting sentences are used.

8. Conclusion:

In this project, an automatic fact verification has been implemented. According to the score of label accuracy and sentence selection, it is ranked at the position of 47 and 32 respectively. The task is split into three steps: (i) To extract the document that contains given claim (ii) Select the most relevant evidence to prove the claim (iii) Tag the label through the analysis of the evidence. Pyluence is the document retrieval tools to locate the document and Word2Vec, Cosine similarity are applied to figure out the best evidence. Finally, it is labeled by the textual entailment model.

References

[1] Tomas, M., Kai, C., Greg, C. Jeffrey, D. Distributed Representations of Words and Phrases and their Compositionality. 1-3. Retrieved from https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[2] Jeffery, P., Richard, S., Christopher D. (2014). GloVe: Global Vectors for Word Representation. *Computer Science Department, Stanford University, Stanford, CA*. 2-4. Retrieved from https://nlp.stanford.edu/pubs/glove.pdf

[3] Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., & Gurevych, I. (2018). UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. *arXiv preprint arXiv:1809.01479*.

[4] Faisal, R., Teruaki, K., Masayoshi, A. (2012) Semantic Cosine Similarity. Graduate School of Science and Technology, Kumamoto University. 1-2. Retrieved from https://pdfs.semanticscholar.org/41ff/3934f40c32ac8643270822de1c763e16c71b.pdf