증거 문장찾기 과업에서 ESIM과 BERT 모델의 한계점 분석

An Application of ESIM and BERT for Evidence Selection Tasks and Result Analyses

Abstract

Automatically detecting fakeness from texts such as news and dialogues is drawing much attention these days, and fact-checking tasks have been proposed to advance the related technology. In these tasks, it is important to give not only the right answer but also evidence that supports it for human consumption. Since the sentence entailment task has a purpose similar to the fact-checking task, we applied two state-of-the-art entailment models to our evidence selection task and analyzed their appropriateness and limitations in the task. While the two models gave reasonable and sufficiently high performance compared to some existing models proposed for a fact-checking competition, we identify some limitations and future work to improve for the evidence selection task.

1. Introduction

With the prevalent fake news, false claims and evidence on the web, fact-checking has become an important task. Given a claim, the task is to automatically verify if it is true or false by identifying related evidence. To verify the truth, finding evidence sentence is essential. Besides, this task is an important yardstick for natural language processing capabilities.

Fact Extraction and VERification (FEVER) [1] is an open challenge to test fact-checking models, which provides a publicly available dataset for fact verification against textual sources. We tested two current state-of-the-art inference and language models on the FEVER dataset to analyze their capabilities and limitations in the evidence selection task. The selected models are Enhanced Sequential Inference Model (ESIM) [2] and Bidirectional Encoder Representations from Transformers (BERT) model [3].

The fact-checking task is considered similar to textual entailment which determines whether a hypothesis is entailed or contradicted by the given premise. In the FEVER 1.0 challenge, most of the models within the top 10 used a modification of the enhanced LSTM model (ESIM) that show state-of-the-art in entailment tasks. BERT [3] is a language model composed of multiple bidirectional Transformer models [4]. It produced a new state-of-the-art performance in many NLP tasks including textual entailment task. We analyze the results of two models so that we verify their capabilities and understand limitations, thereby suggesting future work.

2. Method

FEVER consists of 185,445 claims that can be verified by reading the introductory sections of Wikipedia pages, which are also included in the dataset. A system should classify the claims as Supports, Refutes, or Not Enough Info. If a claim is classified as Supports or Refutes, the system should also return the evidence sentences selected from the retrieved documents, which justify the decision.

Our work concentrates on finding evidence that supports or refutes the claim. The dataset was processed to fit each model and evaluate the models for their respective performances.

As the first step, we retrieve the top 5 document that possibly has an evidence sentence for the given claim based on the retrieval method of [5]. Secondly, all the sentences in the selected documents become candidates for an evidence sentence to be chosen. Golden evidence sentences in the training dataset become positive instances for the classifier and all other sentences in the retrieved documents become a negative instance for the train. Since the number of negative samples is much larger than the positive, we randomly select negative samples for the same number of positive samples.

In the case of ESIM [2], we used modified ESIM which

represents FEVER 1.0 [7]. It gets claim, positive, and negative sentence at the same time to train positive evidence could score higher. In BERT [3], A claim and sentence pair is put into the classifier as the input, and the output label determines whether the sentence is evidence for the claim or not. For the test dataset, we finally obtain a ranking score between two sentences or confidence scores for all the sentences in retrieved documents with respect to the claim. Sentences are ranked by the ranking score and confidence scores respectively.

3. Experiment & Results

We process the train and test dataset to fit the two models. In the case of modified ESIM [7], The model gets positive and negative sentence at the same time, input should be a triple of (claim, positive evidence, negative sentence). A total of 603,131 instances are used for training. For BERT [3], a tuple (claim, sentence) is entered with label 0 (negative sample) or 1 (positive sample). A total of 614,241 instances are used to train BERT.

FEVER 1.0 shows a leader board for top 5 models as in Table 1. Our result beats or shows comparable results compared to this performance in evidence selection. FEVER provides a scoring function for precision, recall, and F1 score [6]. Each model uses best parameter for the number of document retrieval.

Table 1 Top 5 model in FEVER 1.0

Rank	Team Name	Evidence (%)		
		Precision	Recall	F1
1	UNC-NLP	42.27	70.91	52.96
2	UCL Machine Reading Group	22.16	82.84	34.97
3	Athene UKP TU Darmstadt	23.61	**85.19**	36.97
4	Papelo	**92.18**	50.02	**64.85**
5	SWEEPer	18.48	75.39	29.69

Table 2 ESIM, BERT performance for evidence selection. *k* indicates the number of document retrieved.

ESIM (k = 1)	Precision	Recall	F1
Total	0.2434	0.7022	0.3615
Supports	0.2446	0.6956	0.3620
Refutes	0.2416	0.7087	0.3604
BERT (k = 5)	Precision	Recall	F1
Total	0.4696	0.8262	0.5989
Supports	0.4720	0.8203	0.5992
Refutes	0.4636	0.8328	0.5956

4. Error Analysis

ESIM [2] uses local and global inference information by using the interaction between sentences of a pair. BERT [3] uses multiple layers of the transformer model composed of an encoder and a decoder utilizing attention between words. These two models use rich information of interactions between words and sentences based on word embeddings. However, these two models show a typical limitation of embedding-based neural models. Also, attention does not appear to be fit to a certain task.

4.1 ESIM and BERT errors

There are four types of errors. Evidence for the claim has two categories. First, the model should distinguish whether a sentence can be evidence for the claim or not. Second, evidence can support or refute the claim. There are four types of errors: false positive for support, false negative for support, false positive for refute, and false negative for refute. Since BERT shows much higher performance, We analyze the ratio of each error type on BERT to see general tendency.

When we tested on Support and Refute case, BERT shows 91.76% (61,396 / 66,907) accuracy in classifying whether the sentence is evidence or not. Support and Refute case shows little gap in this accuracy (Supports: 93.23%, Refutes: 90.33%). Clearly, making a right decision in Refute cases is harder than Support one. Among total 5515 error cases, four error type shows different ratios.

[supports] false positive: 1200 / 5515 = 21.75%
[supports] false negative: 1045 / 5515 = 18.94%
[refutes] false positive: 1303 / 5515 = 23.62%
[refutes] false negative: 1967 / 5515 = 35.66%

false negative error in finding refute evidence case shows highest error among 4 different type and below section will explain why it is hardest. Below section explain detail examples of wrong answers. To see the qualitative analysis, We analyze 50 samples per error type per model (total 50*4*2 = 400 samples). 50 samples are selected with top 50 highest confidence score in each type to see which kinds of error are confusing most for models. Interestingly, ESIM and BERT shows same category of the error cases. Below sections are major categories of error cases.

4.1.1 Supports/Refutes: False Positives

Omission of Core Parts

Unlike typical information retrieval, finding sentences with overlapping words is not enough for our task. The fact-checking claim has a certain pattern and has a specific attribute to be verified. For example, in 'Mick Thomson was born in Ohio', Both Mick Thomson and Ohio must exist but

with the specific relationship “birth place”. The following is an example of a false positive: 'Mickael Gordon Mick Thomson (born November 3, 1973) is an American heavy metal musician.' It is highly overlapped, but the core part (birthplace) is missing. It shows limitation of attention of ESIM and BERT model; attention learns the word to focus for a high probability of an overlap. They do not attend to a task-specific attribute that could be inferred from the words.

4.1.2 Supports/Refutes: False Negatives

Synonym Problem in words and phrases

Capturing synonym is one of the important factors in sentence pair task. While word-word synonyms are well captured with word embedding, word-phrase synonyms are difficult to deal with. With a claim 'Calcaneal spurs can be detected.' and evidence 'These are also generally visible to the naked eye', it is not easy to match 'can be detected' and 'visible to naked eye' with word embedding based models. First, they must detect the boundary of the phrases. Second, the similarity between the words in the phrase may not be relevant to the similarity between phrases.

Indirect Explanation

In fact-checking, finding lexically similar sentences is not the main purpose. If verification of a claim is possible even with a small part of the sentence, the sentence should be selected as evidence. For example, the claim 'Lucy Hale's middle name is Karen.' has an evidence 'Earlier in her career, she was sometimes credited as Lucy Kate Hale.'. Since the entire sentence is not about the claim proper (only ‘Kate’ is necessary), the sentence is not selected as evidence. This kind of errors is hard to handle using embedding-based approaches. It needs external knowledge about the common sense of word attributes.

4.1.3 Refutes: False Negatives

Same Topic Problems

This is refute-specific error. In evidence for refutes, finding a contradictory word from the claim word is the key for verification. However, a contradicting word usually has the same topic with a very different valence. So high similarity between the claim and sentence cannot be the criterion. For example, 'Paramore is Canadian.' (claim) and 'Paramore is an American rock band from Franklin, Tennessee, formed in 2004.' (evidence) do not have an overlapping word except the main entity 'Paramore'. The attribute of the important word is nationality but since Canadian and American are a bit different, the model generates a false negative. Contradiction in the entire sentence is even more difficult. A claim 'Josh Hutcherson was in a managerial position at a finance company' talks about the career of 'Josh Hutcheson'. The positive evidence 'A native of Kentucky, Hutcherson began his acting career in the early 2000s and appeared in several commercials and minor film and television roles …' is about the career, which is the same topic with the claim, but is entirely contradicted against the claim because it is about the actor career rather than business career.

5. Conclusion

The purpose of textual entailment task is determining whether a hypothesis can be inferred from a premise. A hypothesis can be entailed or contradict the premise. This purpose is similar to fact-checking task of verifying the claim is true or false by selecting the evidence sentences. We show state-of-the-art model in textual entailment can perform well in an evidence selection task and we characterize the failure cases that we have to consider for precise evidence selection models.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. 2013-0-00179, The Development of Core Technology for Context-aware Deep-Symbolic Hybrid Learning and Construction of Language Resources).

Reference

[1] Thorne, James, et al. "FEVER: a large-scale dataset for fact extraction and verification." *arXiv preprint arXiv:1803.05355* (2018).

[2] Chen, Qian, et al. "Enhanced lstm for natural language inference." *arXiv preprint arXiv:1609.06038* (2016).

[3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[4] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. pp. 5998-6008. (2017).

[5] Chakrabarty, Tuhin, Tariq Alhindi, and Smaranda Muresan. "Robust Document Retrieval and Individual Evidence Modeling for Fact Extraction and Verification." *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. pp. 127-131. (2018).

[6] Thorne, James, et al. "The Fact Extraction and VERification (FEVER) Shared Task." *arXiv preprint arXiv:1811.10971* (2018).

[7] Hanselowski, Andreas, et al. "UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification." *arXiv preprint arXiv:1809.01479* (2018).