Improving Large-Scale Fact-Checking using Decomposable Attention Models and Lexical Tagging

Nayeon Lee*, Chien-Sheng Wu*, Pascale Fung

Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Introduction

- Fact-checking of textual sources needs to effectively extract relevant information from large knowledge bases.
- A large-scale fact-checking task, in which verification of claim and extraction of related evidence are required [Thorne et al, 2018]
- Verification labels: Support, Refute and Not enough information (NEI)

Claim	Finding Dory was written by anyone but an American.
Evidence	**Finding_Dory**: Directed by Andrew Stanton with co-direction by Angus MacLane, the screenplay was *written by Stanton and Victoria Strouse* **Andrew_Stanton**: Andrew Stanton -LRB- born December 3, 1965 -RRB- is an *American film director* , screenwriter, producer and voice actor based at Pixar.
Label	REFUTE

Document Retrieval (DR_{rerank})

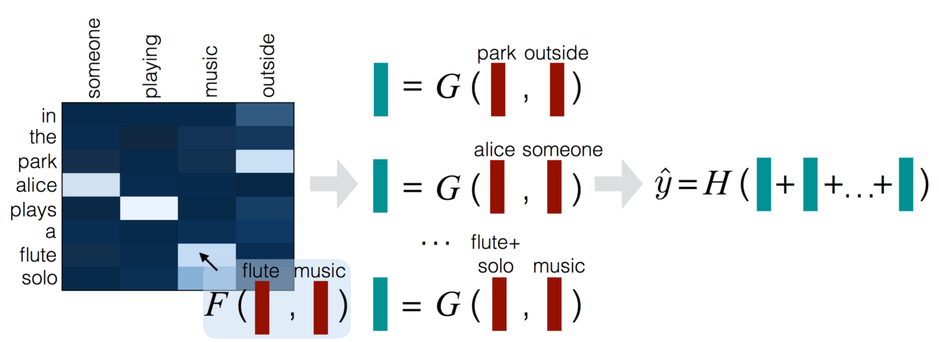
- A document retriever that searches the whole Wikipedia to find the relevant documents
- Use TD-IDF to reduce the search space from 5.4M to 100 documents
- Apply re-ranking using a scoring function f_{rank} that utilizes POS tags (NN, NNS, NNP, NNPS, JJ, CD), then select the top 5 documents.

$$r_{claim} = \frac{\text{POS}_{match}}{\text{POS}_{claim}}, r_{title} = \frac{\text{POS}_{match}}{\text{POS}_{title}},$$
$$f_{rank} = r_{claim} \times r_{title} \times tf\text{-}idf$$

Recognizing Textual Entailment (DA_{rte})

- Given a claim and l possible evidence, a DA_{rte} classifier is trained to recognize the textual entailment to be support, refute or NEI.
- Use the DA between the claim and the evidence for RTE; RTE problem decomposed into sub-problems, which can be considered as bi-direction word-level attention features.

Decomposable Attention (DA)[Parikh et al, 2016]



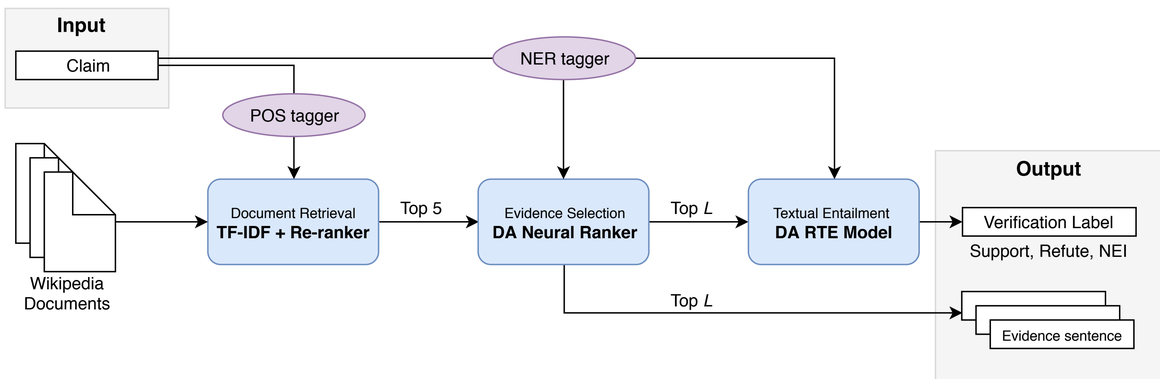
De-noising

th	ES results			RTE results				
	Macro Recall	Macro Precision	F1	Accuracy		Evidence		
				ScoreEv	NoScoreEv	Precision	Recalll	F1
0.2	**0.653**	0.275	0.353	0.405	0.540	0.337	**0.629**	0.439
0.4	0.607	0.349	**0.406**	0.418	**0.542**	0.481	0.586	0.528
0.6	0.535	**0.368**	**0.406**	**0.424**	0.525	0.618	0.517	**0.563**
0.8	0.413	0.330	0.348	0.416	0.484	**0.772**	0.400	0.527

- Prior modules that can effectively leverage the trade-off between recall and precision (high F1) perform the best
- Since the most important factor is to correctly provide succinct set of evidence for the final RTE module.

Proposed architecture

- We propose a framework that verifies a given claim by extracting a set of evidence from Wikipedia.
- We extend an existing pipeline [Thorne et al, 2018] by incorporating lexical tagging and de-noising approaches, and proposing neural ranker.



Evidence Selection (DA_{rank})

- A neural ranker that extracts l sentences as evidence candidates for given claim using decomposable attention (DA) model
- Trained using a fake task, which is to classify whether a given sentence is an evidence of a given claim or not.
- l value is selected dynamically based on the output evidence score of DA_{rank}, which is considered as a confidence measure of a given sentence being an evidence. Evidence with the score below fixed threshold value th is eliminated.

Lexical Tagging

- Part-of-speech (POS) and named entity recognition (NER) are used to enhance the performance.
- Helps in keyword extraction for each claim.
- Reduces the out-of-vocabulary (OOV) problems related to name or organization entities, for better generalization.

Task results

	MLP	DA_{rte}	DA_{rte}+NER
Accuracy (%)	63.2	78.4	**79.9**

Table 1: Oracle RTE classification accuracy in the test set using gold evidence.

		TF-IDF	DA_{rank}			DA_{rank}+NER		
			1:1	1:4	1:9	1:1	1:4	1:9
l	2	0.847	0.170	0.889	0.889	0.109	0.889	**0.893**
	5	0.918	0.451	0.966	**0.968**	0.345	0.962	**0.968**
Time		3.57s	0.055s					

Table 2: Oracle evidence selection macro-recall in the test set using gold documents

Model	Label Accuracy (%)		Label			Evidence F1
	ScoreEv	NoScoreEv	Precision	Recall	F1	
DR_{tfidf} + MLP *	21.80	38.75	0.500	0.387	0.310	0.175
DR_{tfidf} + DA *	30.88	50.44	0.530	0.520	0.517	
Proposed	**42.43**	**52.54**	**0.533**	**0.527**	**0.523**	**0.563**

Table 3: Full-pipeline evaluation on the test set using $k = 2$ and $th = 0.6$.

- Neural ranker allows for faster inference time (×65 speedup) compared to TF-IDF methods that need real-time reconstruction.
- With neural ranker, dynamic evidence selection, we achieve promising improvement in evidence retrieval F1 by 38.80%