Fake News and Fact-checking: Inference and Multilingual Approaches

Proposers: Rodrigo Agerri
Contact: rodrigo.agerri@ehu.eus

Description

One of the most popular tasks and datasets for fake news detection is the Fake News Challenge, an open competition held during 2017 with the aim of investigating how Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies can be leveraged in order to detect and thereby combat fake news in the media. According to the Fake News Challenge: “Fake news, defined by the New York Times as “a made-up story with an intention to deceive” , often for a secondary gain, is arguably one of the most serious challenges facing the news industry today.”[1]

Determining the veracity of a given document or story, namely, whether it is fake or legitimate, is a very complex task, even for expert fact-checkers. Thus, in the Fake News Challenge they decided to break down the fake news detection task in different stages, the first of which is establishing what other news sources are saying about the given document or story (whether they agree, disagree, etc. with the news story), namely, determining their stance with respect to that document or news story. Thus, the first stage of the Fake News Challenge was Stance Detection. This decision was supported by two main ideas:

1. A Stance Detection system should allow a human fact checker to enter a document (headline, message, claim, etc.) and retrieve the top documents from other news sources that agree, disagree or discuss the given document.
2. Based on the previous step, it would be possible to build a “truth-labeling” system based on the weighted credibility of the various news organizations from which the stance has been retrieved.

Figure 1 defines the task of Stance Detection in the Fake News Challenge stage 1. Thus, given a headline and a follow-up news document, the task consists of determining whether the follow-up document agrees, disagrees, discusses, or comments something unrelated with respect to the given news headline. As an example of the task based on this definition, consider the following news snippets:

- **Headline:** “Robert Plant Ripped up $800M Led Zeppelin Reunion Contract.”
- **Follow-up 1 Agrees**: “. . . Led Zeppelin’s Robert Plant turned down £500 MILLION to reform supergroup . . .”

[1] http://www.fakenewschallenge.org/

- **Follow-up 2 Disagrees:** ". . . No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together . . ."
- **Follow-up 3 Discusses**: ". . . Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal . . ."
- **Follow-up 4 Unrelated:** ". . . Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today . . ."

In this project we propose to study Natural Language Inference approaches to detect the stance of a given text with respect to the original trigger headline. Furthermore, we would also like to extend this task to other languages of interest, for which data is very scarce.

Objectives

The candidate may choose between the following objectives:

1. (Semi-) Automatic development of a dataset of fake news detection in languages other than English (for example in the political domain).
2. Experiment with deep learning approaches for fake news, verification and fact-checking, including Natural Language Inference.
3. Reformulate the task as question-answer pairs.

The master thesis can be written in Basque or English.

Tasks and Plan

- Month 1: Start of the project, defining the objectives and tasks.
- Month 2: Start experiments. Optionally, it is recommended for the candidates to attend the "Seminar on language technologies. Deep Learning (LAP 18). https://ixa.si.ehu.es/master/programa_html
- Months 3-5: Experiments and final development.
- Final month: Writing up.

References

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk and Veselin Stoyanov XNLI: Cross-lingual Sentence Understanding through Inference (corpus page) Proceedings of EMNLP, 2018

Conneau et al . (2020). Unsupervised Cross-lingual Representation Learning at Scale (XLM-RoBERTa). In ACL 2020.

Multi-NLI: Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference (corpus page) In NAACL, 2018

Hanselowski et al. (2018) A Retrospective Analysis of the Fake News Challenge Stance Detection Task. In COLING.