News Text in the Hands of a Computational Linguist

Fatemeh Torabi Asr
Discourse Processing Lab

News text has served as a clean, easily accessible and standard type of language data in computational linguistics research. Not only has news corpora helped computational linguists to develop automated tools for parsing, translation and many other general tasks, news text itself has always been an interesting genre of discourse to analyze, among other things, human communication behavior. This talk will cover three research topics dealing in one way or another with news text corpora. I start with introducing my psycholinguistic PhD thesis about discourse relations where the main data was a collection of news articles from Wall Street Journals, i.e., Penn Discourse Treebank. This research was mainly focused on the general usage of discourse connectives: how authors trigger causal, contrastive and other types of relations between discourse segments and whether patterns of connective omission align with cognition and information theory. The second study is related to using large-scale news corpora for socio-political analysis of critical events such as attacks and crises. I will introduce the approach we took to de ne and computationally formalize the challenging concept of events based on participants (named entities) and timestamps (temporal cues) appearing in news text. Finally, I will talk about my current research at the discourse processing lab on automatic fake news detection. Our methodology to tackle missinformation in news is based on the hypothesis that deception has its own linguistic characteristics. Therefore, a fake news article should be objectively di erent from a legitimate news article. We are currently in the process of collecting a large dataset of fake and legitimate news articles. Using suitable discourse analysis and machine learning techniques, we have built a baseline predictive model for automatic fake news detection, which outperforms a random baseline with a signi cant margin. The main concern, however, is to explain the classi cation decisions. I will close the talk with a discussion on this issue, hoping to stimulate great ideas and get feedback for my future research!