Survey on Event Extraction from News Data in Computational Social Science

Aida Mostafazadeh Davani
3893465751

1 News Data in Social Science

The use of news data in different types of analyses is not something new for social scientists and especially political scientist. However, without using Natural Language Processing tools, the data captured by news agencies were manually annotated and structured, before being used by scientist (Hogenboom et al., 2016).

As a result, many individual data collections have been formed over the decades. Some of the most important examples of these type of data collections are: Dimensions of Nations (Rummel, 1972), the Conflict and Peace Data Bank (COPDAB) (Azar, 1980) and its continuation as the Global Event-Data System (GEDS) (Davies and McDaniel, 1993), the Behavioral Correlates of War (BCOW) project (Leng and Singer, 1988), and World Events Interactions Survey (WEIS) (McClelland, 1999).

Formal form of event extraction originates in the late 1980s, when the U.S. Defense Advanced Research Projects Agency (DARPA) boosted research into message understanding, aimed at automating the identification of terrorism-related events from news wires, a topic that has remained trending up until today (Hogenboom et al., 2016).

Ever since this evolution, the advancements of Natural Language Processing techniques have empowered social scientists to exploit “Big Data” for solving research questions. Information extraction approaches enable access to valuable knowledge by retrieving information from unstructured text data.

2 Hate Crime and Morality

The terms hate crime and bias crime were coined in the United States during the 1980s, as journalists and policy advocates groped for new terminology to describe bigoted violence directed against Jews, blacks, and homosexuals (Aronowitz, 1994). Theoretically, hate crime is assumed to be directed by psychological causes, mainly authoritarianism (Altemeyer, 1988) and prejudice (Modena, 2001). More explorations in social psychology suggests that hate is based in group level conflicts (Weisel and Böhm, 2015; Allport et al., 1954).

Based on the unique features of morality, Parker and Janoff-Bulman (2013) suggest that group memberships rooted in moral convictions are a particular classification of inherently dangerous social groups in which outgroup “hate” naturally occurs with ingroup “love”. In this research, we use a specific definition of morality, based on Moral Foundation Theory (MFT). MFT has been shown to be effective in explaining between-group conflicts, including both ideological and violent conflicts (Mooijman et al., 2018).

Mondal et al. (2018) investigates the traces of hate speech in social media (Whisper and Twitter). Due to the lack of location information on Twitter, they couldn’t conclude any information about the geographic attributes of the hate speech. Medina et al. (2018) explores active hate groups in 340 counties of the US and analyzed various political, ideological, and socioeconomic characteristics of those regions. Mondal et al. (2017) studies the online posts in Whisper that include hate speech and how this rhetoric is spread through the US states. Müller and Schwarz (2018) analyzes FBI hate crime reports from different counties of the US and co-occurrence of the rise of hate crimes in a place with Donald Trump’s use of hateful speech on Twitter.

3 Event Extraction

Event extraction is the task of automatically extracting events from unstructured text. This sec-
tion investigates the challenges of event extraction as the primary task that can also include named entity extraction and relation extraction. Recent enhancements in semi-supervised learning and neural networks have improved the accuracy of event extraction tasks vastly.

### 3.1 Semi-supervised Learning

Supervised learning approach has been broadly employed in the context of entity and relation extraction. However, it hugely depends on the specific domain that the task is defined in. The labeled data that is used in supervised learning is always annotated to deal with the requirements of the related field. This fact prevents this approach from being helpful in a more general context. An example of the domain-specific tasks is biological events (Riedel et al., 2011; McClosky et al., 2011).

Other issues that occur in supervised methods of event extraction, especially neural networks, is the sparsity of the data (Estabrooks et al., 2004). The labeled dataset usually contains a small proportion of positive samples. Oversampling methods (duplicating positive data points) and synthetic minority over-sampling (using synthetic data from the minority cluster) or eliminating the negative data points (under-sampling) are the most popular approaches for overcoming this issue (Liu, 2004).

On the other hand, unsupervised methods include clustering and manifold learning of unlabeled datasets and are best for exploring a dataset based on the knowledge we have already captured from previous analyses with similar labeled data. Zhang et al. (2015) introduces an unsupervised algorithm that discovers event relations and then learns to extract them. The algorithm uses a probabilistic graphical model to cluster sentences describing similar events from parallel news streams. These clusters then comprise training data for the extractor.

Semi-supervised learning uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task to overcome the shortcomings of both. Recently, semi-supervised learning mostly finds applications in cognitive psychology as a computational model for human learning since labeling the data that is used in these fields need expert human annotators (Zhu, 2011).

### 3.2 Neural Networks

Elaborately collected featured that are captured from text analysis and are used for recognizing the relations can be categorized into lexical-level and sentence-level clues Chen et al. (2015). Lexical level clues are the features that are extracted to describe words and are mainly supported by prior knowledge. Due to the domain dependence of this type of knowledge, these features are naturally general.

The most challenging aspect of extracting events from a sentence is that the context of a document should be considered in order to interpret an entity and the type of triggered event Chen et al. (2015). Approaches that exclusively use word features for the task, usually lack comprehensiveness. Sentence-level clues are extracted as the context that the relation occurs in to help the event extraction perform more reliably Chen et al. (2015).

Neural networks have been vastly used in event extraction task since they can capture context relations. Chen et al. (2015) represents the use of Dynamic convolutional neural networks (CNN) to apprehend the correct references of words by seizing lexical and sentence level features. The model used in Chen et al. (2015) uses one layer on CNN with max pooling and can extract more than one event from a sentence.

More recent approaches use the recurrent neural network (RNN) to account for the semantic information in a sentence for interpreting the relations. Feng et al. (2018) proposes an unsupervised language independence model for event trigger detection, by exploiting bi-directional long short-term memory (Bi-LSTM) and CNN. The information obtained by the model is proved to be beneficial for sentiment analysis tasks as well as event detection.

Nguyen and Grishman (2015) uses a combination of Bi-LSTM, dependency relations between the words in a sentence, and entity types to represent a sentence. Nguyen and Grishman (2015) compare various combinations of RNNs and CNNs for relation extraction and relation classification and confirm that a majority voting method provides the best result.

### References


(Zhu, 2011).

Nguyen and Grishman (2015) introduces an unsupervised language independence model for event trigger detection, by exploiting bi-directional long short-term memory (Bi-LSTM) and CNN. The information obtained by the model is proved to be beneficial for sentiment analysis tasks as well as event detection.

Nguyen and Grishman (2015) uses a combination of Bi-LSTM, dependency relations between the words in a sentence, and entity types to represent a sentence. Nguyen and Grishman (2015) compare various combinations of RNNs and CNNs for relation extraction and relation classification and confirm that a majority voting method provides the best result.


(Zhu, 2011).


