Towards Large-Scale Event Detection and Extraction from News

Blaz Fortuna, Thomas Demeester, Chris Develder

Ghent University - iMinds, Ghent, Belgium

1 Introduction

Understanding and reasoning about textual data is one of the important topics in artificial intelligence and is being addressed by various research communities, ranging from knowledge representation, over natural language processing to text mining. Each community provides a different set of often overlapping intuitions, tools and methodologies for working with text.

Event processing from news and social media can be seen as a subtopic of text understanding [1, 2, 3, 4]. It comprises different tasks, including New and Retrospective Event Discovery, Event Type Classification and Event Template Extraction.

Research on event processing requires access to annotated data covering different tasks in the event processing pipeline. Over the last decades, several datasets have been created covering event discovery and event extraction, e.g., [5, 6, 7]. These datasets are rather limited in scope. For example, they contain articles from only few selected sources, or they contain a limited number of annotated events with a high selection bias (e.g., towards larger or well defined events like natural disasters or terrorism). Using such limited datasets to evaluate solutions for the event processing tasks may lead to favoring approaches that do not work well on real-world datasets. The main reasons for these limitations are (1) limited access to data resources and (2) the required and expensive manual annotations.

The main contributions we are working towards are (1) a systematic methodology for efficiently creating a large golden standard of manually annotated events over a large corpus of news articles with a realistic distribution over the covered topics and events, and (2) a resulting annotated corpus of 10,000 English general news articles embedded in 31 million news articles.

2 Corpus Selection and Annotation

2.1 Corpus

The corpus is based on a crawl of news articles from Newsfeed [8]. The articles are automatically crawled from RSS feeds of 50,000 news sources, systematically collected to cover news outlets around the world ranging from local to global. Each news source is equipped with metadata describing its title and location. Articles referred in the RSS feeds are crawled and processed in order to extract their title and body. The content of the article is further sent through a language detector, topic classifier and named entity recognizer, and their outputs are attached to the article as metadata.

The corpus contains articles from July 1st, 2013 till June 30th, 2014, totaling 31,015,036 English articles. The corpus will be uniformly sampled for 1000 articles, serving as the seeds for the annotation process.

2.2 Event Definition

There are many ways of defining events. In general events are defined by answering the few simple questions: when, where, who and what. The difference comes from different approaches for event discovery and extraction.

One dimension on which the definitions of events differ is the granularity of text fragments they operate on. On one end of the scale are micro events, which can be extracted from a sentence. In this case, the verb would typically define the event type, and its arguments (e.g., subject, modifiers, object) would define the event parameters. On the other end of the scale are macro events defined by sets of articles, which describe these events. Another dimension is the variation around the definitions of when do two articles talk about the same event. For example, do articles talking about World Cup semi-final match and final match talk about the same event (i.e., "FIFA World Cup 2014"), or two separate events?

In this paper we look at events on the article level, and use the seed articles as definition of events. That is, the seed article provides answers to the questions when, where, who and what.

Article A: FIFA World Cup 2014	
When:	12th June - 13th July, 2014
Where:	Brazil
Who:	All participating teams

Article B: World Cup 2014 final	
When:	13th of July, 2014
Where:	Rio de Janeiro, Brazil
Who:	Germany, Argentina

Figure 1. Two articles defining two different events.

Example: Figure 1 shows the event parameters from two articles; article *A* describes "FIFA World Cup 2014" and article *B* describes "World Cup 2014 final". Each article defines a different event. However, article B provides a partial description of the event defined by article A, and article A can discuss the final in some parts, while not being overall focused on it.

2.3 Annotation Process

The annotation process is a combination of manual annotation and active learning. The process starts afresh from each seed article, to which the annotator manually assigns one or more tags, corresponding to the event type (e.g., "product launch", "sports event"). At this stage the annotator has the option of declaring the article as "non-event" and moving on to the next seed article.

After the seed article is assigned one or more event types, the annotator moves into an active learning phase, with the goal of isolating all the articles related to the event as "positive". In order to initialize the system, we start by sampling random articles according to the cosine similarity (using a traditional bag-of-words representation) with the seed articles. After the initialization we start an active learning loop. In each iteration we select a batch of new articles for human annotation based on uncertainty sampling, in combination with a basic binary text classifier. The stopping criterion for the active learning loop is set on the recall curve measured over the last few annotated batches. Finally, the entire set of articles marked positive are manually annotated as either describing the event (*equal* relation), describing part of event (*part-of* relation), only mentioning the event in some part of the article (*mention* relation), or not related with the event.

As annotators, we used paid students with various backgrounds, working in a close feedback loop to directly clarify potential annotation ambiguities. We also have at our disposal a large living lab with well-known user profiles, ranging from dozens of highly dedicated to multiple thousands of less dedicated users. In the past, we used the smaller groups for more difficult annotation tasks like sentence segmentation, and a larger set of hundreds of users for less intensive tasks such as article labeling. We intend to further employ the living lab for creating our ambitious event detection gold standard.

3 Discussion

The goal of the presented work is to create a large representative golden standard for event discovery and extraction (esp. from general news articles). This corresponds to one of the first steps towards the ultimate goal of automatic extraction and reasoning around events described in news. The corpora will allow for a more realistic development and evaluation of event discovery techniques, which is a mid-term goal of the presented work.

The corpus provides a set of articles and events annotated with event types, which can be further expanded in order to derive a taxonomy of events types. Event types are an important feature when deciding which of the more specific relations are relevant for describing a particular event. For example, events of type “product launch” would be defined by a manufacturer (e.g., Apple), a product (e.g. iPhone) and a date of the launch. On the other hand, events of type “sports event” would have relations such as sports type (e.g., football) and participating individuals/teams.

Extraction of event-based relations is the long-term goal of the presented work. The plan is to extend the annotations from the article level down to the sentence level, and use this annotation to develop models for sentence level extraction of relations. This will allow automatically building and maintaining a knowledge base of events from all around the world, offering a unique and unprecedented view of the global activities.

References

[1] Daniel B. Neill, Weng-Keen Wong. Tutorial on Event Detection, KDD 2009.

[2] Weng, Jianshu, and Bu-Sung Lee. "Event Detection in Twitter." *ICWSM* 11 (2011): 401-408.

[3] Giridhar Kumaran and James Allan. Text classification and named entities for new event detection. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04).

[4] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)

[5] W. Zhao, R. Chen, K. Fan, H. Yan, and X. Li. A novel burst-based text representation model for scalable event detection. In ACL, pages 43–47, 2012.

[6] Wayne, Charles L. "Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation." *LREC*. 2000.

[7] Christopher Walker, Stephanie Strassel, Julie Medero, Kazuaki Maeda. ACE 2005 Multilingual Training Corpus. LDC 2006.

[8] Trampus, Mitja and Novak, Blaz: The Internals Of An Aggregated Web News Feed. Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).