Extracting a Repository of Events and Event References from News Clusters

Silvia Julinda
TU Berlin
Einsteinufer 17
Berlin, Germany
silvia.julinda@gmail.com

Christoph Boden
TU Berlin
Einsteinufer 17
Berlin, Germany
christoph.boden@tu-berlin.de

Alan Akbik
TU Berlin
Einsteinufer 17
Berlin, Germany
alan.akbik@tu-berlin.de

Abstract

In this paper, we prose to build a repository of events and event references from clusters of news articles. We present an automated approach that is based on the hypothesis that if two sentences are *a)* found in the same cluster of news articles and *b)* contain temporal expressions that reference the same point in time, they are likely to refer to the same event. This allows us to group similar sentences together and apply open-domain Information Extraction (OpenIE) methods to extract lists of textual references for each detected event. We outline our proposed approach and present a preliminary evaluation in which we extract events and references from 20 clusters of online news. Our experiments indicate that for the largest part our hypothesis holds true, pointing to a strong potential for applying our approach to building an event repository. We illustrate cases in which our hypothesis fails and discuss ways for addressing sources or errors.

1 Introduction

We present ongoing work in the automatic creation of a repository of events and event references from clusters of online news articles. In the context of this work, an *event* is something that happens at one specific point in time that can be referenced in text with different text surface forms. An example of this may be the acquisition of WhatsApp by Facebook, which has a specific timestamp (02-19-2014), as well as a number of different textual references (such as "the acquisition of WhatsApp", "Facebook's landmark deal" etc). Unlike previous work in event extraction (Aone and Ramos-Santacruz, 2000; Ji and Grishman, 2008), we are less interested in filling slots in a fixed set of event templates. Rather, we aim to identify an unrestricted set of events (Ritter et al., 2012) and all their possible event mentions. This means that even noun phrases ("the much-discussed takeover") and incomplete mentions ("Zuckerberg's 19 billion bet") are valid textual references we wish to capture.

We give examples of such events in Table 1. We believe that automatically distilling such events from news text and hosting them in an event repository could provide a valuable resource to gain a comprehensive overview of world events and also serve as a resource for *event-linking* efforts in future Information Extraction (IE) research.

In this paper, we propose a method for automatically creating such an event repository. Our method leverages computer-generated news sites that aggregate articles from news sources worldwide and group similar stories into news clusters. Such news clusters represent an intriguing reservoir for event extraction: Each cluster typically represents one *news item* that is reported on by hundreds of different online sources. Articles in a cluster will therefore describe similar information content - and reference the same events - using different words. On these news articles, we apply temporal expression taggers to identify and normalize textual references to specific points in time.

Our main hypothesis is that if two sentences are *a)* found in the same cluster of news articles and *b)* contain temporal expressions that reference the same point in time, they are likely to refer to the same event. This allows us to group similar sentences together and for each referenced point in time extract an event with a list of different textual references.

Proceedings of the AHA! Workshop on Information Discovery in Text, pages 14–18,
Dublin, Ireland, August 23 2014.

ID	TIMESTAMP	REPRESENTATIVE	TEXTUAL REFERENCES
1	2014-02-19	Facebook buys WhatsApp	Facebook buying WhatsApp the landmark deal Zuckerberg's acquisition of the mobile messaging-service
2	2014-02-01	Rosetta transmits message	Rosetta sends signal to Earth the spacecraft's first message the message from the Rosetta spacecraft
3	2014-02-07	Sinabung volcano erupts	Indonesian volcano unleashed a major eruption the eruption of Mount Sinabung volcano its biggest eruption yet

Table 1: Examples for events in the event repository. Each extracted event consists of an ID, a timestamp which indicates on which date the event took place, a short human-readable event representation, and a list of strings that may be used to reference this event.

In this paper, we present our event extraction system and conduct a preliminary evaluation in order to determine in how far our hypothesis holds. We discuss the evaluation results and possible improvements and give an outline of current and future work.

2 Event and Reference Extraction

2.1 Method Overview

Determine sentences likely to reference the same event. We begin the event extraction process by crawling Google News[1] to retrieve clusters of English language news articles and their publishing date. Each news article is then boilerplated and segmented into sentences.

We then make use of temporal expression taggers (Strötgen and Gertz, 2010; Chang and Manning, 2012) to recognize temporal expressions in text and normalize them into machine-readable timestamps. This causes expressions such as "last Friday", "winter of 2013", and "Saturday morning" to be normalized to the timestamps "2013-10-10", "2013-WI", and "2012-09-24TMO" respectively by using the article's publishing date as a reference point. We identify all sentences with temporal expressions and group sentences together that *a)* contain the same timestamp and *b)* are found in the same cluster of documents. Refer to Table 2 for examples of sentences grouped according to these criteria.

Determine Open-Domain Facts. Because sentences may refer to multiple events[2], we use OpenIE methods (Schmitz et al., 2012; Choi, 2012) to determine for each sentence a list of N-ary facts. Each fact consist of a predicate and a number of arguments. We then discard all facts that do not contain the temporal expression in order to keep only those facts expressed within each sentence to which the temporal expression refers. This gives us a list of N-ary facts which we presume to refer to the same event, together with its timestamp.

Determine Event Representative and Store. For human readability purposes, we then identify a representative of the grouped N-ary facts by determining the most common predicate and head arguments. We assign a global ID to each event and store it along with its timestamp, its representative and a list of all textual references and their frequency counts in a database.

2.2 Granularity of Timestamps

One question at the onset of this work was which granularity of temporal expressions would be required. We manually inspected a sample of news clusters and noted that news articles rarely provide time information that is accurate to the minute. Rather, most temporal expressions refer to specific dates in past, present or future. We therefore choose the unit "day" as granularity for the temporal expressions in this work. We dismiss all expressions that refer to larger and more vague periods of time ("last winter",

[1] http://news.google.com/

[2] An example of this is the sentence: "*When asked, he said that WhatApp accepted Facebook's offer last Sunday*". Here, the temporal expression "*last Sunday*" refers only the "*WhatApp accepted Facebook's offer*" part of the sentence, not the date the person was asked.

TIMESTAMP	SENTENCES
2014-02-20	Facebook inked a deal **late Wednesday** to buy popular texting service WhatsApp. **Yesterday** , Facebook Chief Executive Officer Mark Zuckerberg bought their five-year-old company. **Thursday , 20 February 2014** Facebook Inc will buy fast-growing mobile-messaging startup WhatsApp. Facebook Inc. agreed to buy mobile messaging-service WhatsApp **today** for as much as 19 billion.
2014-02-01	The European Space Agency received the all-clear message from its Rosetta spacecraft **at 7:18 p.m.** [..] a comet-chasing spacecraft sent its first signal back to Earth **on Monday** ESA received the all-clear message Hello World from its Rosetta spacecraft [..] away **shortly after 7 pm**. **Yesterday's** message from the Rosetta spacecraft was celebrated by scientists [..]
2014-02-07	Indonesia's Mount Sinabung volcano erupted and killed at least 11 people [..] **on Saturday** But **a day later** , Sinabung spewed hot rocks and ash up to 2km in the air. A giant cloud of hot volcanic ash clouds engulfs villages [..] in Sumatra island **on February 1, 2014**. An Indonesian volcano that has been rumbling for months unleashed a major eruption **Saturday**.

Table 2: Examples of sentences grouped by cluster and timestamp. The temporal expression taggers enable us to group sentences that refer to the same point of time in very different ways (highlighted bold). As can be seen in the examples, sentences grouped according to these criteria generally refer to the same event, albeit in sometimes widely varying words.

"throughout the year") and generalize all temporal information that refer to the time of day ("later today", "at 7:18 p.m.").

2.3 Improving Event Quality

Upon manual inspection of identified events we find that our hypothesis fails in some cases: A news item may often summarize a number of smaller events that happened within the same day. An example of this are news items that deal with unrest in war-torn countries that may reference several bombings, terrorist attacks and other violence that happened across the country on the day the article was published. Another example are sports articles that refer to several sport matches that take place during the same day. This is problematic, as in such cases we erroneously link non-synonymous textual references to the same event. We experiment with two methods for reducing this error:

Time Window Filter As indicated above, we note that our hypothesis most often fails for events that occur within 2 days of the publishing date of the articles in the news cluster. Accordingly, we experiment with filtering out such events, leaving only events to be extracted that lie in the more distant past or future (such as past or upcoming election days, significant events that impact the current news story). However, since the largest part of events that are reported on in online news take place within this 2-day time window, we risk significant recall-loss by discarding too many events.

Word Alignment Condition For this reason, we investigate requirements for facts to be grouped together in addition to the requirement of sharing the same timestamp. We experiment with monolingual word-alignment tools (Yao et al., 2013) to determine the "similarity" of two facts as the number of aligned content words. We then require at least one content word to be shared by two facts in order for them to be grouped together into an event.

3 Preliminary Evaluation

We conduct a preliminary evaluation to determine to which extend our hypothesis holds. To this end, we use our method to extract events from a sample of 20 news clusters with an average size of 200 news articles. We evaluate our method in four setups: *1)* The baseline setup in which we apply only the "*same cluster + same timestamp = same event*" hypothesis ("**BASE**"). *2)* The baseline setup plus the time window filter ("**TIME**"). *3)* The baseline setup plus the word alignment condition ("**ALIGN**"). *4)* The baseline setup plus both the time window filter and the word alignment condition ("**ALL**").

We manually evaluate each event found with our method by checking whether all references indeed refer to the same event. We calculate a *purity* value that indicates the proportion of the biggest group

METHOD	TOTAL EVENT REFERENCES	CORRECT EVENT REFERENCES	PRECISION	PURITY
BASE	**609**	511	0.839	0.698
TIME	109	88	0.807	0.699
ALIGN	**609**	**526**	**0.864**	**0.793**
ALL	109	89	0.817	0.728

Table 3: The results of our manual evaluation of extracted events and their event references. The main hypothesis that events mentioned with the same date in one cluster delivers quite promising results with 84% precision. The time window filter does not seem to contribute significant gains, while the ALIGN filter does boost both precision and purity.

of references that refer to the same event over all references in an event cluster. This means that if all references indeed refer to the same event, its purity is *1.0*. Table 3 lists the average purity over all events.

When a reference accurately represents both the content and the date contained in the original news sentence and the real world event mentioned actually occurred on this date, we labeled it as a "correct" event reference. The *precision* listed in Table 3 reflects the proportion of correct events references vs. all extracted event reference in the evaluation data set. This measure indicates how well the extraction itself performs, apart from the clustering of event references.

Hypothesis produces promising results with a precision of 0.84. In general, we find our underlying assumption to indeed be a good basis for event extraction. Our baseline approach based on only this hypothesis produces promising results with a precision of 0.84, albeit at somewhat low overall purity.

Wrong resolution of relative time references biggest source of error. When inspecting sources of errors more closely, we note that the approach fails most often because of erroneous resolution of relative time references such as *"yesterday", "past Saturday"* or *"this Sunday"*. This may happen because the wrong publishing date is assigned to a crawled news article, causing temporal taggers to use a wrong reference point for relative time expressions. With relative references to weekdays, the taggers are often unsure whether the past or present week is referenced. Consider the expression "*on Saturday*" in the sentence "*John Kerry will meet with opposition leaders on Saturday*". Although the coming Saturday is meant in this context, the temporal expression tagger normalizes the date to the last Saturday before the publishing date. We believe that such systematic errors can be addressed in future work through assigning higher confidence to explicit temporal expressions mentions and resolving ambiguities in relative expressions using this information.

Time Window Filter provides no significant contribution. Contrary to initial assumption filtering out events within a 2-day time window does not actually boost precision, but rather greatly reduces the total number of extracted events at slightly lower precision and purity. The likely reason for this behavior is the above noted most common error source is not addressed by this filter.

Word Alignment Condition boosts both precision and purity significantly. The word alignment condition on the other hand greatly increases both precision and purity. While the increase in purity is to be expected as different events occurring on the same date are indeed split into separate clusters, the increase in precision comes as somewhat of a surprise. Closer inspection of the results revealed that the word alignment approach aggressively groups similar event mentions, considering also synonyms as matches, therefore not resulting in redundant event detections as initially feared. Based on these results, we believe that experimentation with word alignment conditions may further increase event detection quality.

4 Conclusion

In this paper, we have proposed to create a repository of events and their textural references and presented an approach to accomplish this automatically by leveraging news clusters and temporal expressions. Our approach is based on the hypothesis that sentences that are found in the same news cluster and refer to the same point in time also refer to the same events. We described the implementation of a prototype system and conducted a preliminary manual evaluation on 20 news clusters to investigate our hypothesis.

Our findings generally point to a strong potential of automatically mining events and references from

news clusters. While our hypothesis fails in some cases, our analysis indicates that incorporating monolingual word-alignment techniques can greatly improve extraction quality and appears to be a powerful tool to disambiguate events that share both timestamp and news cluster.

Present work focuses on further exploring the potential of word alignment as well as the use of cluster-wide statistics to correct labeling mistakes such as the ones observed for temporal tagging. We aim to use the system on very large amounts of news clusters crawled from the Web to generate - and make publicly available - the resource that we have proposed in this paper.

Acknowledgments

The research is funded by the European Union (EU) under grant no. 270137 'SCAPE' in the 7th Framework Program and the German Federal Ministry of Education and Research (BMBF) under grant no. 01ISI2033 'RADAR'.

References

Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *Proceedings of the sixth conference on Applied natural language processing*, pages 76–83. Association for Computational Linguistics.

Angel X Chang and Christopher Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *LREC*, pages 3735–3740.

J. D. Choi. 2012. Optimization of natural language processing components for robustness and scalability.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *ACL*, pages 254–262.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In *Proceedings of ACL*.