A Survey of Event Extraction Methods from Text for Decision Support Systems

Frederik Hogenboom[a], Flavius Frasincar[a,*], Uzay Kaymak[b], Franciska de Jong[c], Emiel Caron[a]

[a] *Erasmus University Rotterdam, PO Box 1738, 3000 DR, Rotterdam, the Netherlands*
[b] *Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, the Netherlands*
[c] *University of Twente, PO Box 217, 7500 AE, Enschede, the Netherlands*

Abstract

Event extraction, a specialized stream of information extraction rooted back into the 1980s, has greatly gained in popularity due to the advent of big data and the developments in the related fields of text mining and natural language processing. However, up to this date, an overview of this particular field remains elusive. Therefore, we give a summarization of event extraction techniques for textual data, distinguishing between data-driven, knowledge-driven, and hybrid methods, and present a qualitative evaluation of these. Moreover, we discuss common decision support applications of event extraction from text corpora. Last, we elaborate on the evaluation of event extraction systems and identify current research issues.

Keywords: Event extraction, information extraction, natural language processing (NLP), text mining

1. Introduction

Over the years, Information Extraction (IE) has become increasingly popular as a tool for a vast array of applications [1–5]. At first, the IE field was focused particularly on message understanding in newswires. However, due to the onset of progressively larger digital data collections of various natural language text types such as news messages, articles, and web pages, researchers and practitioners require more advanced techniques, extract more information with greater accuracies and on a real-time basis, and operate on larger scales than ever before. Since the early 2000's, there has been a notable shift from general information extraction from digital collections – extracting basic named

*Corresponding author; tel: +31 (0)10 408 1340; fax: +31 (0)10 408 9162
Email addresses: fhogenboom@ese.eur.nl (Frederik Hogenboom), frasincar@ese.eur.nl (Flavius Frasincar), u.kaymak@tue.nl (Uzay Kaymak), f.m.g.dejong@utwente.nl (Franciska de Jong), caron@ese.eur.nl (Emiel Caron)

entities like persons and organizations – toward more advanced forms of text mining, including Event Extraction (EE) that requires the handling of textual content or data describing complex relations between entities [6]. This development has been fueled by the continuous advances in Text Mining (TM) and Natural Language Processing (NLP), the advent of big data, as well as the availability of (manually) annotated data sets that often serve as a basis for building extraction models.

Event extraction combines knowledge and experience from a number of domains, including computer science, linguistics, data mining, artificial intelligence, and knowledge modeling. It is commonly seen as the TM-aided extraction of complex combinations of relations between actors (entities), performed after executing a series of initial NLP steps. It is a form of IE, aimed at specific users, applications, and platforms, that results in more complex and detailed outputs than regular IE. 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 newswires, a topic that has remained trending up until today.

With the exponential growth of digital collections and the information extraction requirements in various fields, event extraction research has evolved greatly. Early mentions of modern event extraction can be found in the biomedical literature, where NLP techniques have been traditionally employed for discovering biological entities such as genes and proteins, but where the same techniques are now also widely used for identifying events involving these entities, e.g., gene expressions and protein bindings [7]. Gradually, event extraction has moved to other domains such as politics and finance, where events like elections, CEO changes, or acquisitions, are also comprised of sets of entities (e.g., persons, governments, countries, etc.) and their relations (e.g., leadership, competitor, ownership, etc.) [8, 9].

The detailed information that is commonly extracted from a heterogeneous set of sources in event extraction implementations, becomes increasingly important for supporting decision making processes. Today, the applications of events in decision support systems are plentiful. For instance, events can be used in mediation information systems [10], for the analysis of firm-specific social media monitoring [11], or even for advanced spatio-temporal reasoning in moving objects [12] and vehicle routing [13]. Other popular applications of events lie in environmental scanning [14], news

personalization systems [15], algorithmic trading [16], financial risk analysis [17], e-commerce [18], quality assurance [19, 20], and terrorism detection [21].

Such event-based decision support systems commonly define an event as something that is regarded as happening during a particular interval of time. Events can have multiple occurrences and are generally seen as incidents of substantial importance. In this work, we do not consider organized events such as soccer matches, scientific conferences, parties, etc., but we focus on unexpected occurrences which need to be acted upon. Such events are universally associated with state changes. However, per domain, their definition, complexity, and interpretation could greatly differ.

Irrespective of their domains, extracted events are associated with changes in the state of the current knowledge, and hence can be employed for decision making, prediction, or monitoring. The applications are numerous, e.g., generating trading signals for stock exchange markets, providing event-driven data integration in decision support systems, creating social media monitoring systems by police departments, discovering defects in products, etc. Hence, these developments render traders, managers, and companies to be the users that immediately benefit from event extraction.

Despite the envisaged usefulness and wide prospective applicability of event extraction, several hurdles have to be overcome until event extraction is widely adopted as a supportive tool in practice. The main requirements that were trending in the nineties for information extraction [2], are still applicable to event extraction today. For instance, the technologies should deliver sufficiently accurate results. Furthermore, construction and processing costs should be minimized, and systems are preferred to be operable by non-specialists. These challenging requirements have led to many research efforts in the last decade, of which the main ideas are surveyed in this article.

Although IE in general is certainly a heavily researched and well-described area, to our knowledge, there is little overview work focusing on the upcoming field of event extraction. Therefore, in order to aid researchers and practitioners in making well-informed decisions about their event extraction applications, we survey high-performance extraction techniques and their common applications in decision support systems. While a preliminary survey on event extraction from text already exists [22], here we provide a more complete overview on a higher level of abstraction, and also cover the most recent works. Moreover, in our current endeavors, the various approaches to event extraction are evaluated on more (qualitative) dimensions. We discuss common applications of event

extraction in decision support systems and additionally focus on the evaluation of event extraction methods. Last, current research issues in event extraction from text are highlighted.

2. Techniques

Both in recent research and in practice, a great many of different event extraction techniques have been proposed and applied. In the following discussion on the main techniques that are employed for event extraction, we omit the peculiarities of individual approaches, and focus on several aspects of various commonly applied extraction techniques, identifying their unique properties, advantages, and disadvantages.

A common distinction of event extraction approaches stems from the field of modeling. Data-driven approaches aim to convert data to knowledge through the usage of statistics, data mining, and machine learning. Expert knowledge-driven methods, extract knowledge by exploiting existing expert knowledge, usually through pattern-based approaches. In today's advanced extraction procedures, researchers may employ techniques from both fields by bootstrapping or optimizing their knowledge-based algorithms by means of machine learning, or vice-versa. Those extraction methods that equally employ data- and knowledge-driven techniques can be categorized under the increasingly popular hybrid event extraction approaches.

While preparing this survey, state-of-the-art event extraction articles have been selected based on a thorough examination of the contents of leading journals, such as Decision Support Systems (DSS), ACM Computing Surveys (CSUR), Journal of Biomedical Informatics (JBI), and the like, between the years 2000 and 2015 through the renowned full-text scientific database ScienceDirect, as well as Google Scholar. Initially, we went through the results for the search term '*event extraction*', but after inspection of results for the more general query '*information extraction*', we additionally obtained some other useful related articles. We have subsequently performed a qualitative evaluation of the main types of event extraction techniques[1]. For each method, we analyzed the amount of required data, knowledge, and expertise, as well as the result interpretability, and the required development and execution time. For each of the reviewed works, the criteria were estimated based on information reported in the corresponding papers.

[1]For an overview of event extraction techniques, please see Table A.1 in Appendix A.

Generally, the amount of required data is determined based on reported amounts of documents for which significant results are obtained, where low amounts total to no more than a couple of hundreds of documents, mid-range amounts represent around ten thousand documents, and, in case more documents are required, approaches are considered to be data-intensive. The amount of required knowledge is measured by evaluating the domain specificity of the methods, i.e., the amount of required domain knowledge for executing the necessary extraction steps. This amount is directly proportional to the number of steps requiring domain knowledge, and inversely proportional to the number of commonly applicable methods employed in a single event extraction approach. Hence, a higher number of universal methods lowers the domain specificity, while methods with an emphasis on domain knowledge are more domain specific. The amount of required expertise is determined by analyzing the number of methods that are combined. Also, the (computational) complexity of the methods themselves, the number of required steps, the intricacies of the employed algorithms, etc., exert a notable influence on the amount of required expertise. Result interpretability is observed by evaluating the comprehensibility and traceability of the considered methods. The comprehensibility of the output, i.e., the ease with which results can be translated to human understandable language, is relatively low for numerical outputs, higher for textual results, and maximal for outcomes that incorporate a strong notion of semantics. Black-box methods yield low traceability scores, whereas methods with results that can easily be backtracked (as is the case for – to a lesser extent – grey-box, and – to a wider extent – white-box methods) yield much higher scores. Subsequently, the development time is composed of time invested in (model) construction, training, and parameter tweaking. Last, the execution times (per document) are derived from reported computational complexities.

2.1. Data-Driven Event Extraction

The vast majority of event extraction tools makes use of at least some data-driven techniques, and many of these tools even rely solely on quantitative methods to discover relations. Data-driven approaches develop models of text corpora that approximate linguistic phenomena. Such event extraction techniques are not restricted to basic statistical reasoning based on probability theory, but encompass all quantitative approaches to automated language processing, such as probabilistic modeling, information theory, and linear algebra. These methods focus on specific features, such as

words and n-grams, as well as their associated weights, which are mostly determined using frequency counting algorithms. These features and their associated weights represent the input of complex clustering or classification algorithms, which, despite their differences, all focus on discovering statistical relations, i.e., facts that are supported by statistical evidence. These discovered relations, however, are not necessarily semantically valid, as semantics (meanings) are not explicitly considered, but are assumed to be implicit in the data.

For data-driven event extraction, there is a clear distinction between supervised and unsupervised learning approaches. The former approaches require some expert knowledge, as labeled data is provided to learning algorithms, whereas the latter approaches are usually employed when no labeled data is available. Unsupervised learning is commonly applied in data exploration or structure discovery tasks, and comprises techniques such as clustering and manifold learning. Supervised learning techniques typically produce new events, based on the given labeled examples. Such learning algorithms deduce event properties and characteristics from training data, and use these to generalize to unseen situations. Combining labeled and unlabeled data can produce considerable improvements in learning accuracy, and hence semi-supervised learning methods are often employed when there is a small amount of labeled data, and a large amount of unlabeled data available, for instance when dealing with special, expensive devices or methods.

Popular (supervised) machine learning techniques for learning relations, such as decision trees or neural networks, often prove to be difficult to train for event extraction, due to the fact that these methods require a large amount of data to be trained on, of which much is initially not labeled (annotated). Moreover, the number of negatives (irrelevant data points) tends to largely outweigh the number of positives (relevant data points) in these data sets, which does not only make the number of useful data points sparse, but also adds noise to the trained models. Many techniques exist for tackling this issue of unbalanced data, e.g., over-sampling (duplicating data), under-sampling (removing data), synthetic minority over-sampling (generating synthetic samples), etc. Usually, excessive amounts of negative examples are pruned first from the data, before training extraction models [23].

Another approach to data-driven event extraction is related to inference models, which are very popular in regular IE tasks. These models are mainly used in semi-supervised or unsupervised settings, usually operate on words in sentences or documents, and apply inference on a specific

(learned) probability distribution. The latter distribution is used for predicting the next word in a sentence or document, based on the history of words. For instance, from a corpus it could show that '*ACM*' is frequently followed by '*Press*', but not so much by '*publishing*', yielding a higher probability for '*Press*' to follow '*ACM*' in unseen texts. Inference models use the classification of previous words to predict the next word, by learning which words tend to follow specific words. A commonly used probabilistic model is the n-gram model, where the last word of the n-gram represents the word to be predicted [24, 25].

Clustering of similar or related documents, sentences, terms, etc., is a commonly employed, unsupervised data-driven technique for event extraction. Examples of clustering-driven event extraction approaches are numerous. For instance, one could use clustering on event occurrences over time, and thus predict the type and properties of a new event [26]. Alternative options are clustering documents containing events (parsed through a shallow linguistic analysis) to identify events [27, 28], or sentences referring to the same event [29]. In more complex frameworks, clustering is usually combined with advanced graph structures [30].

The overall trend is that data-driven methods require a lot of data for their training in order to get statistically significant and reliable results. On the other hand, the role for expert knowledge is minimal, as these methods generally do not take into consideration domain semantics, but instead rely on universal, statistical methods that can be applied to any domain. In terms of expertise, however, there is a large variety of data-driven solutions, as the required expertise greatly depends on the methods that are applied. When combining multiple methods, the amount of required expertise is larger than when applying, for instance, merely a single, out-of-the-box clustering method. Also, (semi-)supervised methods generally require more expertise, as labeled data is involved. Although training times are usually long, because of the excessive amounts of data that need to be processed on the one hand, and the computationally intensive operations involved on the other hand, execution times are mostly short for data-driven event extraction methods, as learned weights and parameters are applied to new examples without involving a lot of reasoning on a pre-built model. Last, given the current limitations in the interpretability of machine learning results, we consider these methods to be opaque or semi-transparent at best. Moreover, the interpretability of the results of most data-driven methods is low, because results do not necessarily have explicit semantics associated.

2.2. Knowledge-Driven Event Extraction

Knowledge-driven event extraction methods often use predefined (or learned) patterns expressing expert knowledge rules. Their TM procedures are hence inherently based on linguistic and lexicographic knowledge, as well as on existing human knowledge regarding the content of the texts to be processed. We can make a rough distinction between two types of patterns that can be applied to natural language corpora for event extraction, i.e., lexico-syntactic patterns [31] and lexico-semantic patterns [32]. The former patterns are a combination of lexical representations and syntactic information. The latter patterns are more expressive, and combine lexical representations with both syntactic and semantic information.

Before extraction patterns are employed on a data set of natural language texts, in most knowledge-driven approaches, the corpus is preprocessed using data-driven or knowledge-driven parsers. Most patterns operate on tokens, i.e., small text segments, which are usually words, word groups, numbers, spaces, or punctuation signs. These tokens get assigned various properties, depending on the level of detail and the focus of the NLP pipeline analyzing the corpus. Common properties are a token's associated semantic concept, lexical category, orthographic category, lemma with suffix and/or affix, pronominal reference, etc. Eventually, patterns, constructed according to a predefined grammar, are matched on large collections of tokens. Usually, in case of a match, additional data (properties) are collected and stored in data structures for later usage, e.g., subjects, objects, etc.

When implementing knowledge-driven approaches to perform event extraction tasks, it often proves to be difficult to stay within the boundaries of these approaches, and therefore most methods often have a (small) data-driven component. For instance, initial clustering for classification could be required in order to determine elements that can be used for constructing patterns (e.g., identifying proper nouns, verbs, companies, persons, etc.). In the following discussion on knowledge-driven approaches, we refer to approaches that are fully or mainly pattern-based, as this is the main characteristic of knowledge-driven event extraction methods.

Lexico-syntactic patterns often appear in earlier work on knowledge-driven event extraction [7], but have remained popular in more recent approaches [31, 33] due to their domain independency. The patterns mostly rely on syntactic properties (grammatical meanings) like verbs, nouns, prepositions, and pronouns. Ideally, patterns should be defined in such a way that they occur frequently and thus

cover many event instances. In practice, such patterns cover a limited amount of different statements discussing the same event. Generalizing patterns too much, leads to erroneous matches that fall outside the scope of the intended event, and comes at the cost of a loss in precision, although recall usually increases. As an alternative, one could enumerate all possible verbs and conjugations, but this greatly impacts development times and general rule flexibility. To cope with these and other related issues like synonymy, homonymy, and polysemy, very complex lexico-syntactic patterns need to be constructed. This stresses the need for higher-level patterns, such as lexico-semantic patterns.

Lexico-semantic patterns enable one to extract more accurate information from texts by enriching lexico-syntactic patterns with semantics, i.e., linguistic meaning and domain context. Lexico-semantic patterns allow for more powerful expressions, as they leverage existing lexico-syntactic patterns to a higher abstraction level. In contrast to lexico-syntactic patterns, however, some domain knowledge is required to create high-precision patterns that retrieve many events in an arbitrary corpus. This makes the creation of patterns less trivial, but as the patterns can be less general than lexico-syntactic patterns, they allow for a more specific description of one's needs and return more accurate results at a higher precision level than their lexico-syntactic counterparts, without losing much on the recall level.

In literature, there are two notions of lexico-semantic patterns. Some research is focused on event extraction by means of basic semantics, added through gazetteers (word lists) that are iteratively searched while parsing corpora. As moving from lexico-syntactic approaches to such simple-typed lexico-semantic approaches is a minor incremental step, this approach has often been used, and is for instance exemplified in [17, 32, 34, 35].

Other research aims to use patterns based on ontological concepts and relations, which capture the domain semantics. Ontology-based lexico-semantic patterns involve a more complex typing, as their elements capture domain semantics and are more advanced than syntactic and simple-typed semantic elements. Additionally, restrictions and relations applying to concepts specified in the underlying ontology can be utilized when applying reasoning with an inference engine. Compared to other pattern-based approaches, complex-typed patterns require more expertise due to the increased complexity, yet often generate better results due to their higher expressivity. Most complex-typed lexico-semantic languages, however, reduce their complexity by removing additional features that have been used frequently in lexico-syntactic languages, such as repetition operators or wildcards,

because they focus more on the usage of concepts. Examples of such works are numerous [36–39]. Some languages do offer full expressivity by not only considering ontological classes and relations, but also by additionally supporting labelling, negation, wildcards, and repetition operators [8].

Comparing knowledge-based event extraction methods with data-driven methods generates several insights. In contrast to data-driven methods, knowledge-based techniques require little data, but conversely, they need a considerable amount of linguistic, lexicographic, and – for lexico-semantic patterns – also domain knowledge, inherently increasing the amount of required expertise. Compared to data-driven extraction methods, the emphasis is less on training (development) times, but more on execution times (needed for detecting all the required concepts). Among the knowledge-based methods, there are several differences. Lexico-syntactic patterns require less data for initial training (clustering) phases, as they are closer to the text grammatical structure, yet lexico-semantic patterns often require more development time, due to the need of domain semantics. The results of lexico-semantic patterns additionally excel in interpretability and quality due to their traceability, yet maintenance costs are considerably higher than for lexico-syntactic patterns.

2.3. Hybrid Event Extraction

Research has shown that it is hard to solely apply pattern-based algorithms successfully. These algorithms often need to be bootstrapped or require initial clustering, which can be done by means of statistics. For instance, in [40], the authors apply an initial clustering procedure to their data set of online news articles before acquiring extraction patterns using a semi-supervised machine learning approach. Also, results from knowledge-based approaches are often used in subsequent, data-driven processing steps, e.g., filtering discovered events using term occurrence statistics [41]. Alternatively, methods that traditionally have been statistics-based, such as part-of-speech tagging, can be improved by adding domain knowledge, which is especially useful, for example, in the biomedical domain where common words are often used differently, but also in many other domain-specific and language-specific cases [42]. Moreover, in a similar fashion as presented in [40], knowledge-based event extraction patterns can be learned by applying machine learning techniques, such as conditional random fields, and support vector machines [6, 43–45].

In hybrid event extraction systems, due to the usage of data-driven methods, the amount of required data increases with respect to knowledge-driven systems, yet typically remains less than is

the case with purely data-driven methods. Compared to a knowledge-driven approach, complexity – and hence required expertise – is generally high, as well, due to the combination of multiple techniques. This also often leads to higher training and possibly higher execution times. On the other hand, the amount of expert knowledge required for effective and efficient event discovery is less than for pattern-based methods, because lack of domain knowledge can be compensated by using statistical methods. As for the interpretability, attributing results to specific parts of the event extraction is more difficult due to the addition of data-driven methods. Yet, interpretability still benefits to some extent from the use of semantics as in knowledge-based approaches.

3. Decision Support Applications

The applications of event extraction in decision support systems are very diverse[2], and can be divided into two major fields. First, event extraction has a wide range of utilizations in the biomedical domain [6, 7, 24, 25, 38, 41, 46], for instance for identifying molecular events, protein bindings, and gene expressions, which can subsequently be used in biomedical research. Figure 1 is a typical example of such tools, and depicts the graph-based EVEX user interface for browsing large-scale databases for biomedical events discovered in PubMed abstracts and full-text articles [46]. Here, the nodes represent biological entities like proteins, which are interconnected through edges, representing relations or events. Upon selecting one of these events, many associated properties, such as the type of event, its polarity, the extraction confidence, etc., are retrieved.

Second, many applications of event extraction can be distinguished in news digestion tasks. Usually, event extraction in news is performed for summarization purposes [42] to compress large news messages into a small number of auto-generated sentences based on identified events, but it has also proven to be useful in news personalization systems [15] for selecting relevant news items with respect to user-preferred events. Furthermore, news event applications are found in algorithmic trading [16], and risk analysis [17], where the identified events are often transformed into numerical or binary signals, based on which decisions are made or actions are undertaken.

Most news-oriented event extraction applications are aimed at general news processing [23, 29, 30, 36, 42, 45], but event extraction has also been applied to process scientific [39] and award-

[2]For an overview of event extraction applications, please see Table A.2 in Appendix A.

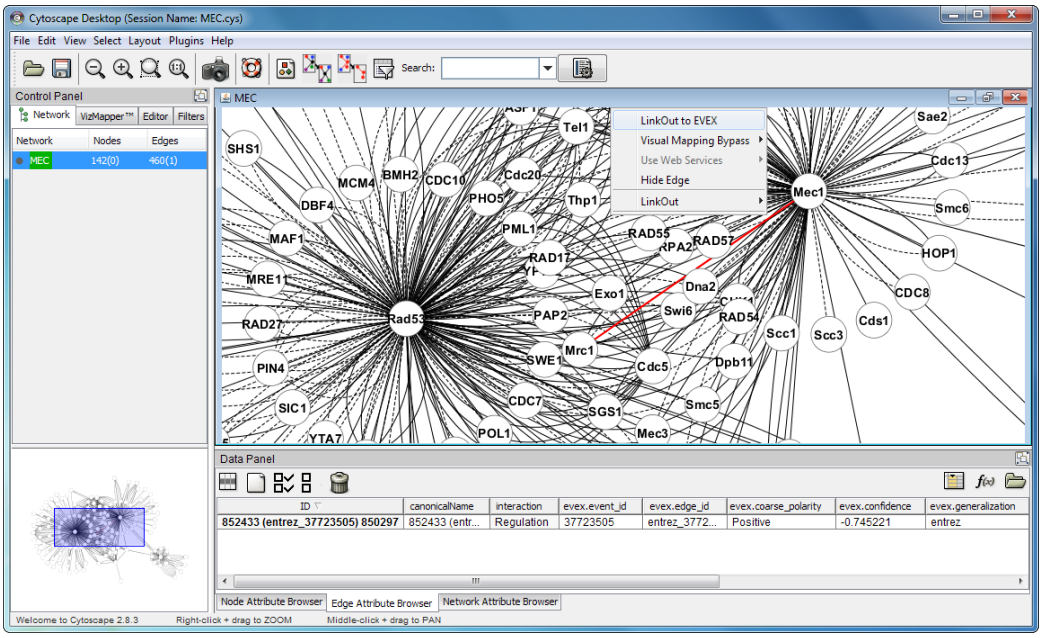

Figure 1: EVEX user interface for browsing large-scale databases for biomedical events.

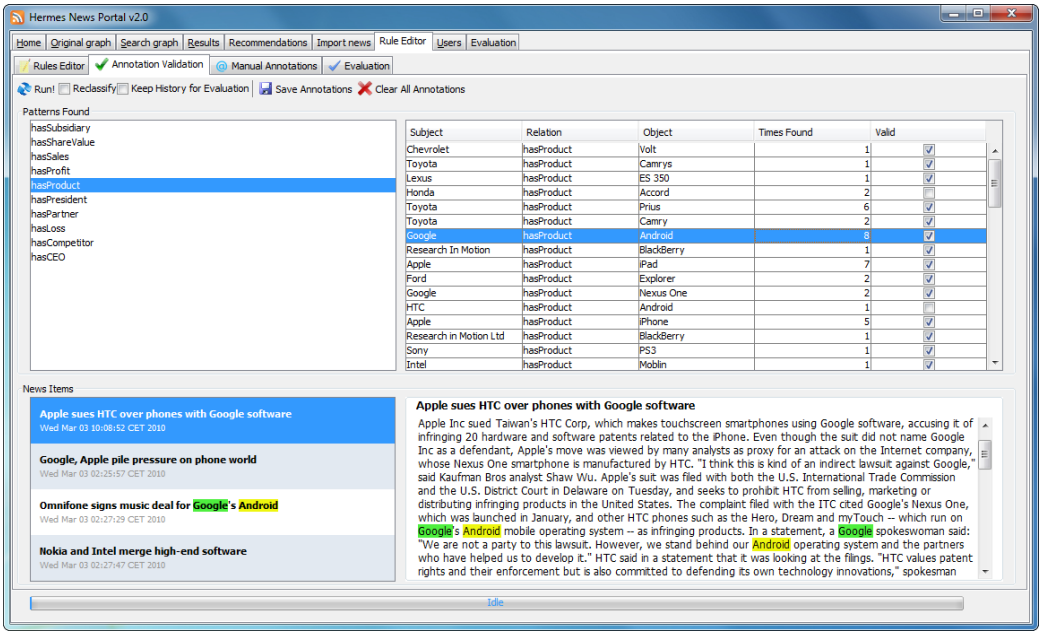

Figure 2: Hermes user interface for browsing news feeds for financial events.

related news [35]. The financial domain is yet another popular application area of news event extraction [8, 32, 37], of which an example, the Hermes News Portal [8], is displayed in Figure 2. Here, financial events are shown to brokers, in order to assist them in daily trading tasks. Events are extracted based on user-defined lexico-semantic patterns, and are displayed to the user for approval. Also, since the 1980s there has been a great demand for event-based solutions for security-related topics such as terrorism, armed conflicts, and epidemiology, which still generates new research outputs today [27, 28, 34, 40].

Other applications found in recent literature are event discovery in legal documents [47], political documents [44], and blogs [26, 33]. Some recent approaches do not limit themselves to written text from documents and streams, but even consider television broadcasts and videos [48] for news summarization and security applications. For this purpose, transcripts are used, but, more recently, research shifted to image processing, e.g., for monitoring systems [49].

4. Evaluation

For the evaluation of event extraction methods, researchers often rely on quantitative indicators, measuring performance using a golden standard-based approach. Data sets, consisting of news messages, documents, articles, etc., are annotated by domain experts, meticulously detailing the events that should be found by the (semi-)automatic event extraction approaches. In accordance with IE and TM, performance is generally measured by computing the number of true positives and negatives, as well as the number of false positives and negatives, each of which can be determined using a golden standard data set, composed by domain experts using a minimum Inter-Annotator Agreement (typically between 60% and 90%, depending on the number of annotators) in order to improve data quality. Based on these numbers, precision (fraction of retrieved events that are relevant), recall (fraction of relevant events that are retrieved), and their harmonic mean, the F_1 score, are computed.

Pre-annotated data sets for event extraction are still rather scarce, as manual annotation is a costly process. The BioNLP'09 shared task on event extraction offers an annotated set[3]. The set is based on the GENIA corpus (a semantically annotated biomedical corpus), and is potentially useful

[3] `http://www.nactem.ac.uk/tsujii/GENIA/SharedTask`

for benchmarking purposes, as 24 teams have reported their final results at the time, and many afterwards. Alternatively, there is a small corpus on general events[4]. Moreover, the workshops on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE), organized in conjunction with the International Semantic Web conferences (ISWC), provide large, general purpose data sets. In the 2011 challenge, a large data set with music and entertainment events is provided[5]. In the 2013 data challenge[6], the focus shifted to linked open data, inviting participants to combine sensor data with maritime data sets on vessels, smuggling, pollution, etc. In the next few years, we expect to see more of such initiatives, resulting in a future defacto benchmarking standard for event extraction, similar to the general Text REtrieval Conference (TREC) IE challenges[7].

The reusability of existing data sets greatly depends on the targeted domains. Therefore, in practice, data are usually scraped from (news) feeds at Reuters, Bloomberg, Yahoo!, etc., after which they are filtered and annotated by domain experts. Crowdsourcing solutions, e.g., services such as Amazon Mechanical Turk[8] and CrowdFlower[9], are great alternatives for obtaining annotations, as they are fast, cheap, and have access to a large pool of potential annotators. Although annotators are usually not domain experts, inconsistencies and inaccuracies can be overcome by using basic measures such as the inter-annotator agreement. Moreover, crowdsourcing services often aid in identifying fraudulent users, and additionally allow for qualification tests in order to further increase annotation quality.

The reusability of existing data sets greatly depends on the targeted domains. Therefore, in practice, data are usually scraped from (news) feeds at Reuters, Bloomberg, Yahoo!, etc., after which they are filtered and annotated by domain experts. Crowdsourcing solutions, e.g., services such as Amazon Mechanical Turk[10] and CrowdFlower[11], are great alternatives for obtaining annotations, as they are fast, cheap, and have access to a large pool of potential annotators. Although annotators are usually not domain experts, inconsistencies and inaccuracies can be overcome by using basic

[4] http://www.isi.edu/~hobbs/EventDuration/annotations/
[5] http://semanticweb.cs.vu.nl/derive2011/Challenge.html
[6] http://derive2013.wordpress.com/data-challenge/
[7] http://trec.nist.gov
[8] http://aws.amazon.com/mturk/
[9] http://crowdflower.com/
[10] http://aws.amazon.com/mturk/
[11] http://crowdflower.com/

measures such as the inter-annotator agreement. Moreover, crowdsourcing services often aid in identifying fraudulent users, and additionally allow for qualification tests in order to further increase annotation quality.

5. Research Issues

In event extraction, there are many open research issues and points of particular interest, of which the main ones are related to: 1) the context-based advantage of data-driven, knowledge-driven, or hybrid approaches, 2) understanding the limitations of specific event extraction techniques, 3) the domain-dependency of event extraction procedures, affecting both their flexibility and effectiveness, 4) the scalability of event extraction approaches when dealing with big data, and 5) the complexity of extracted events.

Similar to what can be observed for the field of IE, there is an ongoing debate on the superiority of data-driven and knowledge-driven approaches to event extraction. Although for both approaches similar performances have been reported in literature in terms of precision, recall, and F_1 scores, advocates of data-driven techniques emphasize their favourable (real-time) computability, whereas knowledge-driven approaches are advocating a higher degree of interpretability due to the general traceability of the results. Users of hybrid event extraction approaches, on the other hand, effectively combine both approaches to their advantage. To determine the best technique for specific applications, there is a need for further research into the best scenarios for the successful application of each technique.

Also, depending on the application at hand, it is important to understand the limitations (and the suitability) of the employed techniques. For instance, opting for knowledge-driven approaches when the quality of annotations cannot be ensured or assessed properly, is generally a bad idea, and data-driven approaches should be considered instead. When the amount of available data is sparse, most data-driven approaches will give results, yet their correctness and reliability is debatable and subject to further evaluation. Furthermore, in the biomedical domain, regular part-of-speech parsers are less useful than in other applications such as (financial) news processing, as there are many special terms, but also common words are used differently and with different meanings, requiring retrained, specialized parsers instead. This is definitely not an isolated case. At the moment, many

researchers acknowledge the existence of such issues in many domains, yet there is little research on the identification of – and principal solutions to – such issues.

Generally, event extraction is a closed-domain procedure. Moving to new domains requires retraining data-driven methods and reformulating patterns used for knowledge-driven event extraction, reducing the overall flexibility of event extraction approaches. Scaling up to larger data sets inherently involves increasing processing power and memory to hold, update, and reason with trained models or knowledge bases. To cope with such problems and with additional issues related to the real-time application of event extraction approaches, research taking into account big data phenomena is necessary. Such research alleviates processing issues associated with large data sets and multi-domain or general knowledge bases, by providing solutions to deal with the scalability problem. For this, researchers aim to optimize algorithms for big data environments (e.g., MapReduce, Spark, etc.).

However, domain and scale changes affect pattern-based approaches in more ways than can be accounted for with solutions borrowed from big data research, requesting additional research issues specifically for knowledge-driven approaches. As the knowledge needed for building patterns is often substantial, such changes drive up the costs for acquiring and maintaining patterns, and additionally increase the danger of consistency errors. Moreover, because humans use natural language in their own unique way, there are many different ways of describing similar information. Patterns have to be highly flexible so that they fit many variants, without compromising their accuracy. Therefore, patterns need to be defined in such a way that they match as many events as possible in an arbitrary corpus (i.e., a high recall), without lowering precision. Up until now, computer-aided pattern learning techniques that aim to alleviate these problems by automating the pattern construction process have not yet resulted in satisfactory results that are comparable to those of hand-crafted patterns. These developments stress the need for additional research into pattern learning and optimization.

Recent event extraction research has primarily focused on extraction procedures and applications of moderate complexity, thereby putting less emphasis on more complex notions of events. Taking into consideration recent developments in sentiment analysis [50], events can be enriched based on the prevailing sentiment regarding the event itself or its (in)directly related actors, associating discovered facts with a sentiment-based weighting scheme so as to increase the utility of events in

for instance decision making processes. Further recent enhancements include the identification of event sequences [51], enabling one to track events over time, and the connection of events to places (e.g., by means of geotags) [52]. However, fully taking into account spatio-temporal aspects while maintaining efficient and accurate reasoning has proven to be a difficult challenge that stimulates further research.

6. Conclusion

Event extraction has recently gained in popularity due to its wide applicability for various purposes. In this article, we reviewed the various data-driven, knowledge-driven, and hybrid techniques of event extraction, and evaluated the works on a set of qualitative dimensions, i.e., the amount of required data, knowledge, and expertise, as well as the interpretability of the results and the required development and execution times. We identified the major strengths and weaknesses of the main event extraction techniques, as well as their major differences. Data-driven approaches require a lot of data available and little domain expertise, while knowledge-driven approaches work adequately on small data sets, but require more expert knowledge. Hybrid methods inherit the benefits of both data-driven and knowledge-driven approaches, mitigating their disadvantages. Moreover, we discussed the main application areas of event extraction, e.g., the biomedical, security, financial, quality assurance domains, etc. In addition, we provided a discussion on the evaluation of event extraction systems. Last, we identified several research issues that need to be addressed, such as approach scalability and domain dependencies.

In the near future, we envisage event extraction to evolve in various ways. First, current encouraging developments in sentiment analysis [50, 53] can stimulate event extraction research by connecting sentiment to events, which are currently often merely rich facts decorated with actors and other properties. Also, as the current field is already moving toward identifying event sequences [51], tracking events over time, and connecting events to places (e.g., by means of geotags) [52], a next possible step in event extraction could be to fully take into account spatio-temporal aspects, not only by connecting these aspects to events, but also by exploiting this information in reasoning and discovering new events. Further event enrichment with domain properties is also a promising research direction, e.g., determining the related obligations and permissions [18], the associated revenues, costs, profits, importance, priority, or criticality, or even the specific entities affected by

events [20]. Furthermore, we envisage better real-time performances due to improved hardware and the rise of computing clusters, but also due to the output of current and ongoing research into big data, resulting in scalable solutions. Last, with the advent of linked open data, large accessible knowledge bases such as DBpedia, and linked semantic lexicons [54], many more application domains can be supported, reducing the need for creating and maintaining gazetteering lists and ontologies for knowledge-based event extraction techniques (open event extraction). The latter developments will make event extraction more accessible and trustworthy, facilitating the development of previously unenvisaged applications of event extraction.

Acknowledgment

The authors are partially supported by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and the Dutch national program COMMIT.

References

[1] D. E. Appelt, Introduction to Information Extraction, AI Communications 12 (3) (1999) 161–172.

[2] J. Cowie, W. Lehnert, Information Extraction, Communications of the ACM 39 (1) (1996) 80–91.

[3] O. Etzioni, M. Banko, S. Soderland, D. S. Weld, Open Information Extraction from the Web, Communications of the ACM 51 (12) (2008) 68–74.

[4] R. Grishman, Information Extraction A Multidisciplinary Approach to an Emerging Information Technology, Springer, 1997, Ch. Information Extraction: Techniques and Challenges, pp. 10–27.

[5] M.-F. Moens, Information Extraction: Algorithms and Prospects in a Retrieval Context, Vol. 21 of The Information Retrieval Series, Springer, 2006.

[6] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, T. Salakoski, Complex Event Extraction at PubMed Scale, Bioinformatics 26 (12) (2010) i382–i390.

[7] A. Yakushiji, Y. Tateisi, Y. Miyao, Event Extraction from Biomedical Papers using a Full Parser, in: 6th Pacific Symposium on Biocomputing (PSB 2001), 2001, pp. 408–419.

[8] W. IJntema, J. Sangers, F. Hogenboom, F. Frasincar, A Lexico-Semantic Pattern Language for Learning Ontology Instances from Text, Journal of Web Semantics: Science, Services and Agents on the World Wide Web 15 (1) (2012) 37–50.

[9] Z. Ma, O. R. Sheng, G. Pant, Discovering Company Revenue Relations from News: A Network Approach, Decision Support Systems 47 (4) (2009) 408–414.

[10] A.-M. Barthe-Delanoë, S. Truptil, F. Bénaben, H. Pingaud, Event-Driven Agility of Interoperability during the Run-Time of Collaborative Processes, Decision Support Systems 59 (2014) 171–179.

[11] S. Jiang, H. Chen, J. F. Nunamaker, D. Zimbra, Analyzing Firm-Specific Social Media and Market: A Stakeholder-Based Event Analysis Framework, Decision Support Systems 67 (2014) 30–39.

[12] B. Jin, W. Zhuo, J. Hu, H. Chen, Y. Yang, Specifiying and Detecting Spatio-Temporal Events in the Internet of Things, Decision Support Systems 55 (1) (2013) 256–269.

[13] V. Pillac, C. Guéret, A. L. Medaglia, An Event-Driven Optimization Framework for Dynamic Vehicle Routing, Decision Support Systems 54 (1) (2012) 414–423.

[14] C.-P. Wei, Y.-H. Lee, Event detection from Online News Documents for Supporting Environmental Scanning, Decision Support Systems 36 (4) (2004) 385–401.

[15] J. Borsje, F. Hogenboom, F. Frasincar, Semi-Automatic Financial Events Discovery Based on Lexico-Semantic Patterns, International Journal of Web Engineering and Technology 6 (2) (2010) 115–140.

[16] W. Nuij, V. Milea, F. Hogenboom, F. Frasincar, U. Kaymak, An Automated Framework for Incorporating News into Stock Trading Strategies, IEEE Transactions on Knowledge and Data Engineering 26 (4) (2014) 823–835.

[17] P. Capet, T. Delavallade, T. Nakamura, A. Sandor, C. Tarsitano, S. Voyatzi, Intelligent Information Processing IV, Vol. 288 of IFIP International Federation for Information Processing, Springer, 2008, Ch. A Risk Assessment System with Automatic Extraction of Event Types, pp. 220–229.

[18] A. S. Abrahams, Developing and Executing Electronic Commerce Applications with Occurences, Doctoral Dissertation, University of Cambridge (2002).

[19] A. S. Abrahams, J. Jiao, G. A. Wang, W. Fan, Vehicle Defect Discovery from Social Media, Decision Support Systems 54 (1) (2012) 87–97.

[20] A. S. Abrahams, J. Jiao, W. Fan, G. A. Wang, Z. Zhang, What's Buzzing in the Blizzard of Buzz? Automotive Component Isolation in Social Media Postings, Decision Support Systems 55 (4) (2013) 871–882.

[21] S. J. Conlon, A. S. Abrahams, L. L. Simmons, Terrorism Information Extraction from Online Reports, Journal of Computer Information Systems 55 (3) (2015) 20–28.

[22] F. Hogenboom, F. Frasincar, U. Kaymak, F. de Jong, An Overview of Event Extraction from Text, in: ISWC Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), Vol. 779 of CEUR Workshop Proceedings, CEUR-WS.org, 2011, pp. 48–57.

[23] Z. Lei, L.-D. Wu, Y. Zhang, Y.-C. Liu, A System for Detecting and Tracking Internet News Event, in: Advances in Multimedia Information Processing, Vol. 3767 of LNCS, Springer, 2005, pp. 754–764.

[24] S. Riedel, H.-W. Chun, T. Takagi, J. Tsujii, A Markov Logic Approach to Bio-Molecular Event Extraction, in: Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task (BioNLP 2009), Association of Computational Linguistics, 2009, pp. 41–49.

[25] M. Miwa, R. Sætre, J.-D. Kim, J. Tsjujii, Event Extraction With Complex Event Classification Using Rich Features, Journal of Bioinformatics and Computational Biology 8 (1) (2010) 131–146.

[26] M. Okamoto, M. Kikuchi, Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries, in: Information Retrieval Technology, Vol. 5839 of LNCS, Springer, 2009, pp. 181–192.

[27] M. Atkinson, J. Piskorski, H. Tanev, E. van der Goot, R. Yangarber, V. Zavarella, Automated Event Extraction in the Domain of Border Security, in: User Centric Media, Vol. 40 of Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer, 2009, pp. 321–326.

[28] H. Tanev, J. Piskorski, M. Atkinson, Real-Time News Event Extraction for Global Crisis Monitoring, in: Natural Language and Information Systems, Vol. 5039 of LNCS, Springer, 2008, pp. 207–218.

[29] M. Naughton, N. Kushmerick, J. Carthy, Event Extraction from Heterogeneous News Sources, in: AAAI Workshop on Event Extraction and Synthesis (W8), From: `http://www.aaai.org/Papers/Workshops/2006/WS-06-07/WS06-07-002.pdf`, 2006.

[30] M. Liu, Y. Liu, L. Xiang, X. Chen, Q. Yang, Extracting Key Entities and Significant Events from Online Daily News, in: Intelligent Data Engineering and Automated Learning, Vol. 5326 of LNCS, Springer, 2008, pp. 201–209.

[31] S.-H. Hung, C.-H. Lin, J.-S. Hong, Web Mining for Event-Based Commonsense Knowledge Using Lexico-Syntactic Pattern Matching and Semantic Role Labeling, Expert Systems with Applications 37 (1) (2010) 341–347.

[32] F. Li, H. Sheng, D. Zhang, Event Pattern Discovery from the Stock Market Bulletin, in: Discovery Science, Vol. 2534 of LNCS, Springer, 2002, pp. 35–49.

[33] Y. Nishihara, K. Sato, W. Sunayama, Event Extraction and Visualization for Obtaining Personal Experiences from Blogs, in: Human Interface and the Management of Information. Information and Interaction. Part II, Vol. 5618 of LNCS, Springer, 2009, pp. 315–324.

[34] M. Atkinson, M. Du, J. Piskorski, H. Tanev, R. Yangarber, V. Zavarella, Computational Linguistics, Vol. 458 of Studies in Computational Intelligence, Springer, 2013, Ch. Techniques for Multilingual Security-Related Event Extraction from Online News, pp. 163–186.

[35] F. Xu, H. Uszkoreit, H. Li, Automatic Event and Relation Detection with Seeds of Varying Complexity, in: AAAI Workshop on Event Extraction and Synthesis (W8), From: `http://www.aaai.org/Papers/Workshops/2006/WS-06-07/WS06-07-004.pdf`, 2006.

[36] C. Aone, M. Ramos-Santacruz, REES: A Large-Scale Relation and Event Extraction System, in: 6th Applied Natural Language Processing Conference (ANLP 2000), ACL, 2000, pp. 76–83.

[37] E. Arendarenko, T. Kakkonen, Ontology-Based Information and Event Extraction for Business Intelligence, in: Artificial Intelligence: Methodology, Systems, and Applications, Vol. 7557 of LNCS, Springer, 2012, pp. 89–102.

[38] K. B. Cohen, K. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner, Jr., E. White, H. Tipney, L. Hunter, High-Precision Biological Event Extraction with a Concept Recognizer, in: NAACL-HLT Workshop on BioNLP: Shared Task, ACL, 2009, pp. 50–58.

[39] M. Vargas-Vera, D. Celjuska, Event Recognition on News Stories and Semi-Automatic Population of an Ontology, in: 3rd IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), IEEE Computer Society, 2004, pp. 615–618.

[40] J. Piskorski, H. Tanev, P. O. Wennerberg, Extracting Violent Events From On-Line News for Ontology Population, in: Business Information Systems, Vol. 4439 of LNCS, Springer, 2007, pp. 287–300.

[41] H.-W. Chun, Y.-S. Hwang, H.-C. Rim, Unsupervised Event Extraction from Biomedical Literature Using Co-occurrence Information and Basic Patterns, in: Natural Language Processing, Vol. 3248 of LNCS, Springer, 2004, pp. 777–786.

[42] C.-S. Lee, Y.-J. Chen, Z.-W. Jian, Ontology-Based Fuzzy Event Extraction Agent for Chinese E-News Summarization, Expert Systems with Applications 25 (3) (2003) 431–447.

[43] C. Best, J. Piskorski, B. Pouliquen, R. Steinberger, H. Tanev, Intelligence and Security Informatics, Vol. 135 of Studies in Computational Intelligence, Springer, 2008, Ch. Automating Event Extraction for the Security Domain, pp. 17–43.

[44] F. Jungermann, K. Morik, Enhanced Services for Targeted Information Retrieval by Event Extraction and Data Mining, in: Natural Language and Information Systems, Vol. 5039 of LNCS, Springer, 2008, pp. 335–336.

[45] M.-V. Tran, M.-H. Nguyen, S.-Q. Nguyen, M.-T. Nguyen, X.-H. Phan, VnLoc: A Real – Time News Event Extraction Framework for Vietnamese, in: 4th International Conference on Knowledge and Systems Engineering (KSE 2012), IEEE Computer Society, 2012, pp. 161–166.

[46] S. van Landeghem, J. Björne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. van de Peer, F. Ginter, Large-Scale Event Extraction from Literature with Multi-Level Gene Normalization, PLoS One 8 (4) (2013) e55814.

[47] N. Lagos, F. Segond, S. Castellani, J. O'Neill, Event Extraction for Legal Case Building and Reasoning, in: Intelligent Information Processing V, Vol. 340 of IFIP Advances in Information and Communication Technology, Springer, 2010, pp. 92–101.

[48] M. Chen, C. Zhang, S.-C. Chen, Semantic Event Extraction Using Neural Network Ensembles, in: 1st IEEE International Conference on Semantic Computing (ICSC 2007), Computer Society, 2007, pp. 575–580.

[49] S. Kamijo, Y. Matsushita, K. Ikeuchi, M. Sakauchi, Traffic monitoring and accident detection at intersections, IEEE Transactions on Intelligent Transportation Systems 1 (2) (2000) 108–118.

[50] R. Feldman, Techniques and Applications for Sentiment Analysis, Communications of the ACM 56 (4) (2013) 82–89.

[51] C. K. Kengne, L. C. Fopa, A. Termier, N. Ibrahim, M.-C. Rousset, T. Washio, M. Santana, Efficiently Rewriting Large Multimedia Application Execution Traces with Few Event Sequences, in: 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013), ACM, 2013, pp. 1348–1356.

[52] S.-S. Ho, M. Lieberman, P. Wang, H. Sarnet, Mining Future Spatiotemporal Events and their Sentiment from Online News Articles for Location-Aware Recommendation System, in: 1st ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems (MobiGIS 2012), ACM, 2012, pp. 25–32.

[53] A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, F. de Jong, Multi-Lingual Support for Lexicon-Based Sentiment Analysis Guided by Semantics, Decision Support Systems 62 (2014) 43–53.

[54] L. Shi, R. Mihalcea, Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing, in: Computational Linguistics and Intelligent Text Processing, Vol. 3406 of LNCS, Springer, 2005, pp. 100–111.

Appendix A. Summary of Surveyed Works

The tables in this appendix summarize the surveyed works. Table A.1 shows the event extraction techniques, which are grouped by their class (distinguishing between data-driven, knowledge-driven, and hybrid methods), and subsequently ordered by their appearance. Our findings with respect to the amount of required data, knowledge, expertise, development time, execution time, as well as the interpretability of the results can be found in their respective columns, along with a short description of the techniques elaborated on in the referred work. Table A.2 provides an overview of the discussed event extraction applications. We keep track of the source type and application domain(s), and give a short description of the reviewed application. The items are grouped by their source and domains, and are subsequently ordered according to their appearance.

Ref.	Class	Data	Knowledge	Expertise	Dev. Time	Ex. Time	Interpretability	Description
[23]	Data	Large corpus	Domain independent methods	Several steps, high complexity, labeled data	Slow supervised learning	Fast parameterized model	Black-box method	Data pruning before training Support Vector Machines
[24]	Data	Medium corpus	Domain independent methods	Few steps, medium complexity, labeled data	Slow semi-supervised learning	Fast probabilistic model	Black-box method, clean relational structures output	Joint probabilistic modeling and Markov Logic inference
[25]	Data	Medium corpus	Domain independent methods	Few steps, high complexity (rich features), labeled data	Slow semi-supervised learning	Fast parameterized model	Black-box method	Machine learning using rich (complex) features
[26]	Data	Medium corpus	Domain independent methods	Few steps, very high complexity	Slow unsupervised learning	Fast distance computation	Black-box method	Two-level hierarchical clustering
[27]	Data	Large corpus	Shallow linguistics, geotagging	Several steps, medium complexity, data already preclassified	Slow unsupervised learning	Fast distance computation	Black-box method	Clustering based on content similarity

Continued on next page

Continued from previous page

Ref.	Class	Data	Knowledge	Expertise	Dev. Time	Ex. Time	Interpretability	Description
[28]	Data	Large corpus	Shallow linguistics, geotagging, manually correcting learned patterns	Several steps, high complexity, data already pre-classified	Slow unsupervised learning, manual corrections	Fast distance computation and parameterized model	Black-box method, understandable patterns output	Clustering based on content similarity, machine learning
[29]	Data	Large corpus	Domain independent methods	Few steps, very high complexity	Slow unsupervised learning	Fast distance computation	Black-box method	Hierarchical agglomerative clustering of sentences
[30]	Data	Large corpus	Domain independent methods	Several steps, very high complexity	Slow unsupervised learning	Fast distance computation	Black-box method, graph structure is insightful	Clustering and advanced graph structures
[7]	Knowledge	Medium corpus	Domain independent methods require linguistic knowledge	Substantial amount of processing steps, high complexity	Long creation times for complex patterns	Slow syntactic parsing	White-box method, basic lexico-syntactic patterns	Pattern matching on canonical structure
[8]	Knowledge	Small corpus	Domain dependent methods require linguistic and financial knowledge	Large amount of processing steps, high complexity	Concept-aided, expressive pattern creation is less time consuming	Slow semantic parsing and ontological reasoning	White-box method, expressive and intuitive complex-typed lexico-semantic patterns	Ontology-based event extraction using highly expressive lexico-semantic patterns and ontological reasoning

Continued on next page

Continued from previous page

Ref.	Class	Data	Knowledge	Expertise	Dev. Time	Ex. Time	Interpretability	Description
[17]	Knowledge	Small corpus	Domain dependent methods require linguistic and financial knowledge	Large amount of processing steps, high complexity	Pattern creation is time consuming	Slow semantic parsing, fast logics	White-box method, verbose simple-typed lexico-semantic patterns	Pattern-based extraction of events for usage in fuzzy logic warning system
[31]	Knowledge	Small corpus	Domain independent methods require linguistic knowledge	Substantial amount of processing steps, high complexity	Manually crafting sets of complex patterns is time consuming	Slow syntactic parsing	White-box method, intuitive lexico-syntactic patterns	Text matching in web search results using designed lexico-syntactic patterns
[32]	Knowledge	Medium corpus	Domain dependent methods require linguistic and financial knowledge	Substantial amount of processing steps, high complexity	Some training involved, no manual pattern construction	Slow semantic parsing	Grey-box method, simple-typed lexico-semantic patterns	Chinese lexico-semantic event discovery by linking keywords to concepts
[33]	Knowledge	Small corpus	Domain independent methods require linguistic knowledge	Substantial amount of processing steps, high complexity	Low amount of training involved, no manual pattern construction	Implicit pattern construction and blog retrieval slows down computation	Grey-box method, unclear lexico-syntactic patterns, difficult picture/text output interpretation	Event construction using collected keywords from past-tense sentences in blogs

Continued on next page

Continued from previous page

Ref.	Class	Data	Knowledge	Expertise	Dev. Time	Ex. Time	Interpretability	Description
[34]	Knowledge	Small corpus	Domain dependent patterns require linguistic and security knowledge	Large amount of processing steps, high complexity	Pattern creation is time consuming due to verbosity and complexity	Slow semantic parsing	White-box method, simple-typed lexico-semantic patterns, clear output	Domain-specific lexico-semantic pattern construction for event extraction
[35]	Knowledge	Medium corpus	Domain dependent methods require linguistic and scientific news knowledge	Substantial amount of processing steps, high complexity	Pattern learning is time consuming	Slow semantic parsing	White-box method, simple-typed lexico-semantic patterns	Pattern learning using syntactic and basic semantic concepts
[36]	Knowledge	Small corpus	Domain dependent patterns require linguistic and general knowledge	Substantial amount of processing steps, high complexity, ontology	Concept-aided pattern creation is less time consuming	Slow semantic parsing	White-box method, complex-typed lexico-semantic patterns	Large-scale pattern-based end-to-end event extraction
[37]	Knowledge	Small corpus	Domain dependent patterns require linguistic and business knowledge	Substantial amount of processing steps, high complexity	Pattern creation in JAPE is time consuming due to verbosity	Slow semantic parsing	White-box method, verbose complex-typed lexico-semantic patterns	Ontology-based event extraction system

Continued on next page

Continued from previous page

Ref.	Class	Data	Knowledge	Expertise	Dev. Time	Ex. Time	Interpretability	Description
[38]	Knowledge	Medium corpus	Domain dependent methods require linguistic and biomedical knowledge	Substantial amount of processing steps, high complexity	Concept-aided pattern creation is less time consuming	Slow semantic parsing	White-box method, accessible complex-typed lexico-semantic patterns	Event extraction using manually constructed patterns exploiting ontological concepts
[39]	Knowledge	Small corpus	Domain dependent methods require linguistic and scientific news knowledge	Substantial amount of processing steps, high complexity	Concept-aided pattern creation is less time consuming	Slow semantic parsing	White-box method, basic complex-typed lexico-semantic patterns	Ontology population through event extraction using lexico-semantic patterns
[6]	Hybrid	Large corpus	Domain dependent methods require some biomedical knowledge	Large amount of processing steps, high complexity	Slow supervised learning	Slow semantic parsing, fast parameterized model	Black-box method	Classifying unknown sentences with Support Vector Machines using graph-based semantic representations
[40]	Hybrid	Large corpus	Domain dependent methods require some security knowledge	Large amount of processing steps, high complexity	Slow semi-supervised learning	Slow semantic parsing, fast parameterized model	Grey-box method, patterns understandable	Bootstrapping weakly-supervised pattern acquisition of clustered news

Continued on next page

Continued from previous page

Ref.	Class	Data	Knowledge	Expertise	Dev. Time	Ex. Time	Interpretability	Description
[41]	Hybrid	Large corpus	Domain dependent methods require some biomedical knowledge	Large amount of processing steps, high complexity	Slow unsupervised learning	Slow semantic parsing, fast parameterized model	Black-box method	Filtering discovered events using term occurrence statistics, pattern learning
[42]	Hybrid	Medium corpus	Domain dependent methods require some meteorological knowledge	Large amount of processing steps, high complexity	Slow supervised learning	Slow semantic parsing, fast parameterized model	Grey-box method, ontology concepts clear	Ontology-based data-driven parsing of Chinese weather messages
[43]	Hybrid	Large corpus	Domain dependent methods require linguistic and security knowledge	Large amount of processing steps, high complexity	Slow semi-supervised learning	Slow semantic parsing, fast parameterized model	Grey-box method, machine learning opaque	Automating event extraction by learning patterns
[44]	Hybrid	Large corpus	Domain dependent methods require some political knowledge	Large amount of processing steps, high complexity	Slow supervised learning	Slow semantic parsing, fast parameterized model	Black-box method	Using undirected graphical models to improve named entity recognition (disambiguation)

Continued on next page

Continued from previous page

Ref.	Class	Data	Knowledge	Expertise	Dev. Time	Ex. Time	Interpretability	Description
[45]	Hybrid	Medium corpus	Domain dependent methods require some security knowledge	Large amount of processing steps, high complexity	Slow semi-supervised learning	Slow semantic parsing, fast parameterized model	Grey-box method, machine learning opaque	Learning lexico-semantic extraction patterns

Table A.1: Overview of event extraction techniques.

Ref.	Source	Domain(s)	Description
[6]	Articles	Biomedics	PubMed-driven knowledge base generation for gene expressions, regulations, phosphorylations, etc.
[7]	Articles	Biomedics	MEDLINE abstract parsing using general purpose tools
[24]	Articles	Biomedics	Collection of bio-molecular events in BioNLP tasks
[25]	Articles	Biomedics	Collection of bio-molecular events in BioNLP tasks
[38]	Articles	Biomedics	Collection of bio-molecular events in BioNLP tasks
[41]	Articles	Biomedics	Aggregation of events from the GENIA corpus with MEDLINE abstracts
[46]	Articles	Biomedics	Extraction of gene and protein events from PubMed articles (BioNLP / BioCreative tasks)
[26]	Blogs	General	Identification of volatile events and places of interest from various locations in Tokyo
[33]	Blogs	General	Collection and visualization of personal experiences
[47]	Documents	Legal	Extraction of events for legal case building using lawyers' documents on key players
[44]	Documents	Politics	Compilation of decision making events in the German parliament's plenary sessions and petitions
[35]	E-News	Awards, Science	Identification of scientific (Nobel, Pulitzer, etc.) prize-winning events

Continued on next page

Continued from previous page

Ref.	Source	Domain(s)	Description
[15]	E-News	Finance	News personalization using extracted financial events
[16]	E-News	Finance	Algorithmic trading using event-based signals
[17]	E-News	Finance	Adapted financial risk analysis based on financial events
[8]	E-News	Finance, Politics	Notification of financial and political events to support stock market traders
[23]	E-News	General	Tracking news stories over time
[30]	E-News	General	Identification of key entities and topics in Chinese news
[36]	E-News	General	Selection of relations of places, organizations, and persons from various sources
[39]	E-News	Science	Acquisition of events related to academic life at KMi
[27]	E-News	Security	Multi-lingual extraction of border security events
[28]	E-News	Security	Extraction of violent events from Europe Media Monitor news articles
[34]	E-News	Security	Multi-lingual identification of security-related events in Europe Media Monitor news articles
[40]	E-News	Security	Population of ontologies using extracted violent events
[45]	E-News	Security	Extraction of accidents, crimes, disasters, etc. from Vietnamese news (BAOMOI)
[29]	E-News	War	Extraction of war-related events from multiple sources
[42]	E-News	Weather	Chinese news summarization
[49]	Images	Security	Traffic monitoring and accident detection at intersections
[48]	Videos	Sports	Summarization of soccer matches

Table A.2: Overview of event extraction applications.