A Survey of Textual Event Extraction from Social Networks

Mohamed MEJRI[1] and Jalel AKAICHI[1]

Institut Supérieur de Gestion de Tunis

Abstract. In the last decade, mining textual content on social networks to extract relevant data and useful knowledge is becoming an omnipresent task. One common application of text mining is Event Extraction, which is considered as a complex task divided into multiple sub-tasks of varying difficulty. In this paper, we present a survey of the main existing text mining techniques which are used for many different event extraction aims. First, we present the main *data-driven* approaches which are based on statistics models to convert data to knowledge. Second, we expose the knowledge-driven approaches which are based on expert knowledge to extract knowledge usually by means of pattern-based approaches. Then we present the main existing hybrid approaches that combines data-driven and data-knowledge approaches. We end this paper with a comparative study that recapitulates the main features of each presented method.

Key-words: Event Extraction, Text Mining, Information Extraction, Social Network.

1 Introduction

Social Networks are defined as web-based systems (dedicated websites or other application) that allow users (individuals) to create public or semi-public profiles and communicate with each other, within the internet network, by posting information, comments, messages, videos, etc. [4, 8].

In recent years, Social networks have become omnipresent because of the increasing propagation and affordability of internet enabled devices such as personal computers, smart phones, tablets and many other devices that allow users to connect to social networks through the internet services [3]. These new possibilities allow people from everywhere and anytime to add, update, share and consult massive quantities of new information in real time. These huge quantities of new information added by hundreds of millions of active users [26] are considered as a very important source of data for many research fields.

These massive quantities of data are characterized by three computational issues: size, noise and dynamism [4]. These issues make manual analysis of social network data seems to be impossible. To remedy this problem, data mining provides a wide range of techniques for detecting useful knowledge from massive datasets. Most of data social network is initially unstructured and habitually described using human natural language, which makes the understanding and interpretation of social network content by machine a difficult task [6]. This problem impedes the automation of Text Mining (TM) sub-tasks such as Information Retrieval [13]and Information Extraction (IE) [15] processes which are frequently used in the decision making.

In general, we can define Text Mining (TM) as the analysis of data contained in natural language text. TM works by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques [40]. By means of text mining, often using Natural Language Processing (NLP) techniques, information is extracted from texts of various sources, such as news messages and blogs, and is represented and stored in a structured way, (generally in databases). A specific type of knowledge that can be extracted from text by means of TM is an event, which can be represented as a complex combination of relations linked to a set of empirical observations from texts [17]. Event Extraction (EE) from textual content of social network has gained remarkable attention in the last few years. For example, this representation *<person>* *<attack>* *<person>* presents *an attack event.* Words identified in text referring to persons are linked to the concept *<person>*; verbs having the meaning of attack are associated with *<attack>*. Thus, a similar event representation can be detected from texts such as: "John shot his friend", "A woman was attacked by a stranger." Etc.

Saval et al. [33] proposed a semantic extension for the modeling of events type "natural disasters". They define an event **E** as the combination of three components: a semantic property **S**, a time interval I, and a spatial entity SP. Thus, an event is represented as follows: **E < I; SP; S>**. In their work of 2014, Serrano et al. [34] adapted this event representation by enriching it with an additional component **A** corresponding to the different participants involved in the event. Thus, the representation will be as follows: *<I, SP, S, A>* where **A** is a set of participants playing one or more role(s). **A** participant noted P_i wherein $0 < i < n$, and a role noted r_j wherein $0 < j < k$. Component A is then defined as follows: $A = \{(P\alpha, r\beta)\}$ as the participant $P\alpha$ plays the role $r\beta$ in the concerned event.
Event extraction from unstructured textual content could be useful for IE systems in various ways. In fact, being able to detect and recuperate events could enhance the quality and performance of personalized systems [14]. Therefore, the use of extracted events form textual content of social networks to deal with several issues is becoming an unavoidable task. However, Extracting events is a very difficult task divided to many sub-tasks with different complexities and need the combination of many techniques and methods depending on the treaty task.
In this paper, we present a survey of the main existing approaches in literature for EE. In the first section, we present the data-driven event extraction approaches, which are based on methods relying on statistics to convert data to knowledge, then, we expose the main knowledge-driven approaches which extract knowledge through representation and exploitation of expert knowledge, usually by means of pattern-based approaches. The last part of the first section will be devoted to the presentation of different hybrid methods based on the combination of data-driven and knowledge-driven approaches. In section 3, we present a quick overview of the main multilingual event extraction systems used in the recent literature. In the third section, we discuss the main existing works that combine event extraction and risk management. And we end this papers with a comparative study in which we demonstrate the main differences, advantages and disadvantages for each approach.

2 Event extraction from textual content

In the available annotated corpora geared toward information extraction, we see two models of events, emphasizing these different aspects. On the one hand, there is the TimeML model, in which an event is a word that points to a node in a network of temporal relations. On the other hand, there is the ACE model, in which an event is a complex structure, relating arguments that are themselves complex structures, but with only ancillary temporal information (in the form of temporal arguments, which are only noted when explicitly given). In the TimeML model, every event is annotated, because every event takes part in the temporal network. In the ACE model, only "interesting" events (events that fall into one of 34 predefined categories) are annotated. The task of automatically extracting ACE events is more complex than extracting TimeML events (in line with the increased complexity of ACE events), involving detection of event anchors, assignment of an array of attributes, identification of arguments and assignment of roles, and determination of event coreference.

Events in the ACE program

The ACE program1 provides annotated data, evaluation tools, and periodic evaluation exercises for a variety of information extraction tasks. There are five basic kinds of extraction targets supported by ACE: entities, times, values, relations, and events. The ACE tasks for 2005 are more fully described in [1].

ACE entities fall into seven types (person, organization, location, geo-political entity, facility, vehicle, weapon), each with a number of subtypes. Within the ACE program, a distinction is made between entities and entity mentions (similarly between event and event mentions, and so on). An entity mention is a referring expression in text (a name, pronoun, or other noun phrase) that refers to something of an appropriate type. An entity, then, is either the actual referent, in the world, of an entity mention or the cluster of entity mentions in a text that refer to the same actual entity. The ACE Entity Detection and Recognition task requires both the identification of expressions in text that refer to entities (i.e., entity mentions) and coreference resolution to determine which entity mentions refer to the same entities.

ACE events, like ACE entities, are restricted to a range of types. Thus, not all events in a text are annotated—only those of an appropriate type. The eight event types (with subtypes in parentheses) are Life (Be-Born, Marry, Divorce, Injure, Die), Movement (Transport), Transaction (Transfer-Ownership, Transfer-Money), Business (Start-Org, Merge-Org, Declare-Bankruptcy, EndOrg), Conflict (Attack, Demonstrate), Contact (Meet, Phone-Write), Personnel (Start-Position, End-Position, Nominate, Elect), Justice (ArrestJail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon). Since there is nothing inherent in the task that requires the two levels of type and subtype, for the remainder of the paper, we will refer to the combination of event type and subtype (e.g., Life:Die) as the event type. In addition to their type, events have four other attributes (possible values in parentheses): modality (Asserted, Other), polarity (Positive, Negative), genericity (Specific, Generic), tense (Past, Present, Future, Unspecified).

The most distinctive characteristic of events (unlike entities, times, and values, but like relations) is that they have arguments. Each event type has a set of possible argument roles, which may be filled by entities, values, or times. In all, there are 35 role types, although no single event can have all 35 roles. A complete description of which roles go with which event types can be found in the annotation guidelines for ACE events [38]. Events, like entities, are distinguished from their mentions in text. An event mention is a span of text (an extent, usually a sentence) with a distinguished anchor (the word that "most clearly expresses [an event's] occurrence" [38]) and zero or more arguments, which are entity mentions, timexes, or values in the extent. An event is either an actual event, in the world, or a cluster of event mentions that refer to the same actual event. Note that the arguments of an event are the entities, times, and values corresponding to the entity mentions, timexes, and values that are arguments of the event mentions that make up the event. The official evaluation metric of the ACE program is ACE value, a cost-based metric which associates a normalized, weighted cost to system errors and subtracts that cost from a maximum score of 100%. For events, the associated costs are largely determined by the costs of the arguments, so that errors in entity, timex, and value recognition are multiplied in event ACE value. Since it is useful to evaluate the performance of event detection and recognition independently of the recognition of entities, times, and values, the ACE program includes diagnostic tasks, in which partial ground truth information is provided. Of particular interest here is the diagnostic task for event detection and recognition, in which ground truth entities, values, and times are provided.

According to ACE terminology, *event trigger* is the word that determines the event occurrence; *argument* is an entity mention, a value or a temporal expression that constitutes event attributes and *event mention* is an extent of text with the distinguished trigger, entity mentions and other argument types [15].
As mentioned above, event extraction is a complex task divided on many subtasks; therefore, many techniques for event extraction from textual content exist in literature. As will be shown in this paper, the choice of suitable techniques is based on the final requirements of each extraction task. In this section, we present a survey on the main methods and approaches sued in recent literature: the data-driven approaches, knowledge-driven approaches and the hybrid approaches, we end this section by a comparative study that recapitulating the main features, fields of application, advantages and disadvantages of each approach.

2.1 Data-driven approaches for event extraction

In contrast to pattern-based approaches (which are presented in section 2.2), data-driven approaches automatically build models for a particular NLP tasks (i.e. to automated language processing) with no human intervention. In other words, these approaches try to discover statistical relations through the use of only quantitative methods such as probabilistic modeling, information theory, and linear algebra. So, to develop these models that approximate linguistic phenomena, data-driven methods necessitate a large text corpora, which is why these techniques often are called corpus-based. Examples of discovered facts are words or concepts that are (statistically) associated with one another. In recent literature, many techniques associated to data-driven approaches could be used such as: word frequency counting, Term Frequency - Inverse Document Frequency (TF-IDF), word sense disambiguation (WSD), n-grams, and clustering.

One common task in data-driven approaches for event extraction from text is the Part-of-Speech (POS) tagging which is the process of assigning a part-of-speech to each word in a sentence. In their work of 2006, Guy et al [11] elaborated on a comparison between four data-driven taggers (TnT, MBT, SVMTool and MXPOST). The experiments obtained through the application of these data-driven taggers on a given dataset (the annotated Helsinki Corpus of Swahili) shows that MXPOST as being the most accurate tagger for this dataset. In another set of experiments, they further improved on the performance of the individual taggers by combining them into a committee of taggers. Likewise, the obtained results showed that combining many taggers may enhance the performance and accuracy of system. In the same field and to deal with for morphologically complex languages,, Mark et Joel [12] extended a statistical tagger to handle fine grained tagsets and improve over the best Icelandic POS tagger. Additionally, they develop a case tagger for non-local case and gender decisions. Delia et al. [31] investigated different unsupervised techniques for extracting and clustering complex events from news articles. As a first step they proposed two complementary event extraction algorithms, based on identifying verbs and their arguments and shortest paths between entities, respectively. Next, they obtained more general representations of the event mentions by annotating the event trigger and arguments with concepts from knowledge bases. The generalized arguments were used as features for a clustering approach, thus determining related events.
In their work of 2014, Deyu et al [41] elaborated on a simple Bayesian modeling approach to event extraction from Twitter, called Latent Event Model (LEM), to extract structured representation of events from social media. However, the proposed

approach is fully unsupervised and does not require annotated data for training. So, the proposed model only requires the identification of named entities, locations and time expressions. After that, the model can automatically extract events which involving a named entity at certain time, location, and with event-related keywords based on the co-occurrence patterns of the event elements. Okamoto et al. [27] presented a method for the detection of occasional or volatile local events using topic extraction technologies. They elaborate on a framework based on a two-level hierarchical clustering method. The resort to clustering techniques gave acceptable results with a good accuracy for event extraction. Liu et al. [5] presented a framework for simultaneous key entities extraction and significant events mining from daily web news based on clustering, modeling entities and weighted undirected bipartite graph. In the same filed, the authors of [37] developed a real-time news event extraction system based on automatic pattern learning from a small annotated corpus and in order to guarantee that massive amounts of textual data can be digested in real time, they have developed ExPRESS (Extraction Pattern Engine and Specification Suite), a highly efficient extraction pattern engine, which is capable of matching thousands of patterns within seconds. In [24], Lei et al. presented a framework for extracting and tracking topic relevant event based on SVM algorithm.

The use data-driven approaches for event extraction give a main advantage: there is no need to expert knowledge or linguistic resources. However, data-driven approaches require large text corpora in order to develop models that approximate linguistic phenomena. Another drawback is that data-driven methods do not deal with the meaning of text. To remedy this problem, researchers resort to knowledge-driven approaches which are based on patterns that express rules representing expert knowledge.

2.2 Knowledge-driven approaches for event extraction

Also known as Rule-Based methods, knowledge-driven methods are commonly based on patterns constructed by linguists. Patterns consist of lexically specified syntactic templates that are matched to text, in much the same way as regular expressions, which are applied along with type constraints on substrings of the match. These patterns are lexically indexed local grammar fragments, annotated with semantic relations between the various arguments and the knowledge representation [39]. So, these rules or patterns are relying on linguistic knowledge about the structure of language and written in a formal notation so that they used by the computer for further parsing [25]. The design of patterns (that may be lexico-syntactic or lexico-semantic pattern) and the choose of appropriate techniques are generally depends on many factors such as the language of the text that is to be processed and the final purpose of processing. For the lexico-syntactic case, patterns combine lexical and syntactical information [22] while for the case of lexico-semantic patterns are employed by the addition of semantic information generally through the use of gazetteers [19] or ontologies [20].

Lexico-syntactic patterns

As we mention before, lexico-syntactic patterns is a combinations between lexical representations (i.e., strings) and syntactical information (e.g., Part-Of-Speech). For further clarification, we present the following lexico-syntactic pattern given by Hearst in his work of 1998 [16]:

such NP as {NP,}$*$ {(or | and)} NP

Where he aimed to find hyponym and hypernym relations by discovering regular expression patterns in free text. In this pattern, “NP” indicates a proper noun. Other text (i.e., “such”, “as”, “or”, and “and”) is used for lexical matching, while “(” and “)” contain conjunction and disjunction statements to be evaluated, in this case a disjunction (denoted as “|”). Also, “∗” is a repetition parameter that indicates the sequence between braces (“” and “”) is allowed to repeat zero to an infinite number of times. Apply this lexico-syntactic pattern on this sentence “... works by such authors as Herrick, Goldsmith, and Shakespeare” gives the following results:
hyponym("author", "Herrick")
hyponym("author", "Goldsmith")
hyponym("author", "Shakespeare")

These patterns are often easy to comprehend by regular users, yet defining the right patterns to mine corpora to obtain unknown information is not a trivial task. Hearst stresses that, in order to return desired results successfully, patterns should be defined in such a way that they occur frequently and in many text genres. Also, they should often indicate the relation of interest and should be recognizable with little or no pre-encoded knowledge. Furthermore, all existing syntactic variations have to be included into a complex pattern to ensure its proper working.

Lexico-semantic patterns

Lexico-semantic patterns are employed to remedy problems of the absence of concepts that have specific meaning (mean by the use of lexico-syntactic patterns). In addition to the combination of lexical representations and syntactical information used by lexico-syntactic patterns, lexico-semantic patterns also permit for the usage of semantic information such as concepts that are defined in ontologies. So, Lexico-semantic patterns combine lexical representations with syntactic and semantic information. Lexico-semantic patterns are first presented by [21] in their work of 1991, where they made a system for text processing based on lexico-semantic patterns. These patterns could include terms and operators like lexical features, logical combinations, and repetition, which are mostly adopted from the regular expression language.
The following example is given by Wooter el al [7] is a lexico-semantic pattern that will classify the verb phrase “left dead” as to express death or injury:
?PIVOT = (or found left shot)
?OBJ =∗ ?EFFECT=dead
=> (mark-activator
murder d-vp) ;

This sentence would also match “found dead” and “shot dead”. Next to standard elements such as repetition and wildcards, the rule presented here contains features like variable assignment on the left-hand side (LHS) (where words preceded by “?” denote variables) and on the right-hand side (RHS) macros such as “mark-activator”, which uses the results of the pattern match, including variable assignments, along with some other constants, such as “murder” and “d-vp”, to tag and segment the text. The use of lexico-semantic patterns gives many advantages, the most important is that they take into account the domain semantics which help the parser cope with the complexity and flexibility of unstructured text written with natural language [16].

In the current body of literature, many works based on knowledge-driven approaches for event extraction exists. For instance, in their work of 2012, Wooter et al [19] proposed a rule-based method to learn ontology instances from text, where

they defined a lexico-semantic pattern language that, in addition to the lexical and syntactical information present in lexico-syntactic rules, also makes use of semantic information.
In [16], authors proposed the use of lexico-semantic patterns for extracting financial events from RSS news feeds in order to allow investors on financial markets to monitor financial events when deciding on buying and selling equities. These patterns use financial ontologies, leveraging the commonly used lexico-syntactic patterns to a higher abstraction level, thus enabling lexico-semantic patterns to recognize increasingly precisely events than lexico-syntactic patterns from text. For that, authors have developed rules based on lexico-semantic patterns used to find events, and semantic actions that allow for updating the domain ontology with the effects of the discovered events. There, pattern creation was based on the triple paradigm (i.e., it makes use of a subject, a predicate, and an optional object), and that relies on triple conversion to the Java Annotations Pattern Engine[1] (JAPE) language [10] and SPARQL[2] [2]. Another work for economic event extraction is also presented for the same authors [18], in which they proposed a semantic-based information extraction pipeline for economic event detection, which makes use of lexico-semantic patterns that are defined in the JAPE language. Other works in the same field could be found in [35], [36].

The resort to knowledge-driven approaches has alleviated many problems figured in case of data-driven approaches. The first issue fixed by the employ of knowledge-driven approaches is that we don't need to use a huge amount of training data (text corpora demanded by data-driven approaches) to develop models that approximate linguistic phenomena. The second important advantage is that the remedy to knowledge-driven approaches offers the possibility to rely on a combination of lexical, syntactical and semantic elements to define powerful patterns which can be used to extract and recognize very specific information. Nevertheless, one common negative point concerns knowledge-driven approaches is that prior domain knowledge is required, so we need to ask for expert linguist help, in other words, , in order to be able to define patterns that retrieve the correct, desired information, lexical knowledge and possibly also prior domain knowledge is required. Also, the resort only to knowledge-driven approaches may cause troubles and returns weak results especially when we need to recognize a big number of events.

2.3 Hybrid approaches for event extraction

Staying within the limits of one type of event extraction approaches may not give the best results. So, combining data-driven approaches with knowledge-driven ones possibly will alleviate drawbacks of each kind and this actually creates a new kind of approaches: the hybrid approaches. In practice, it's difficult to rely only on one kind of event extraction approaches. Therefore, the majority of works in the recent literature relies on hybrid approaches. Generally, and during the application of hybrid approaches, data-driven approaches are generally used for the statistical processing (bootstrapping, POS tagging, initial clustering, etc) while knowledge-driven approaches are used for defining powerful expressions generally by means of lexical, syntactical and semantic elements [29]. In other words, data-driven approaches used to deal with huge amount of data while knowledge-driven used to deal with specific meaning aims.
Kenji et al. [32] presented an approach to combine rule-based and data-driven NLP

[1] https://gate.ac.uk/sale/tao/splitch8.html
[2] http://www.w3.org/TR/rdf-sparql-query/

techniques in the extraction of grammatical relations. They have shown that starting with a rule-based system, we can use unlabeled data and a corpus-based system to improve recall (and F-score) of grammatical relations. In their work of 2004, Camiano et al. [9] elaborated on a hybrid approach to resolve issues caused by the lack of expert knowledge, so they resort to statistical methods to remedy these issues. Pakhomov et al. [28] combined statistical methods with lexical knowledge. A similar orientation could be found in [30] in this case, authors used hybrid approach to reinforce statistical methods. The authors of [29] bootstrap a weakly supervised pattern learning algorithm with clusters, in order to extract violence incidents from online news with high precision and recall, and storing these in knowledge bases. The authors of [23] employ a grammar-based statistical method to text mining, i.e., POS tagging. However, tagging is based on domain knowledge that is stored in ontologies, thus making the event extraction a hybrid process. Finally, Chun et al. [15] extract events from biomedical literature by means of lexico-syntactic patterns, combined with term co-occurrences.

The combination of data-driven approaches with knowledge-driven ones bring several enhancements. For instance, and even still need a big amount of data to develop statistical models, the required amount of data in hybrid approaches is less than in the case of purely data-driven approaches. The same, the required amount of developed patterns by experts for detecting events is less than purely knowledge-driven approaches and this is due to the resort to statistical methods to discover events automatically. Drawbacks are generally caused by the complexity of hybrid systems which encompasses many techniques and methods of data-driven and data-knowledge approaches.

3 Discussion

In this section, we summarized the different discussed approaches and methods in a table (Table 1), in which we tried to expose the main differences between each approach. To do so, we listed, the different techniques used for each approach (Data-driven or knowledge driven approaches) then the used methods for each approach (hierarchical, graphs, SVM. . .) and the different types of events. We presented, also, the amount of required data needed for each approach and finally the required domain knowledge and expertise). As shown in Table 1, in term of data usage, knowledge driven based approaches require fewer amounts of data. Experiments shows that we need only couple hundreds of documents or sentences to generate valuable and accurate event extraction rules. On the other hand, data-driven approaches require more than ten thousands documents to build useful statistical models that give acceptable results. For the hybrid approaches that combine data-driven and knowledge-driven methods, the amount of required data still elevated but it's much better than the case of Data-driven approaches, where we rely solely on statistical techniques to extract rules. For the interpretability, knowledge-driven approaches give the best results, especially for the case of lexico-semantic patterns that performs the high level of interpretability. The data-driven approaches give the lowest accurate. Based on the results given by this survey, and in order to chose the appropriate techniques and methods for event extraction, we recommend the resort to knowledge-driven approaches for specific domains, due the ease, the simplicity and the high accurate of rules based approaches. Also we need less amount of data to generate useful models. In the other hand, we recommend data-driven and hybrid approaches for users who deal with huge amount and variety of data to extract various types of events.

Table 1. A comparison between the 3 event extraction categories in terms of: amount of necessary data, demanded knowledge and expertise

Technique	Approach	Method	Events	Data	Knowledge	Expertise	Interpretability
Data	Guy et al [11]		data-driven taggers	Med	Low	Low	Low
	Mark et Joel [12]			High	Low	Low	Low
	Delia et al. [31]			Med	Low	Low	Low
	Deyu et al [41]			High	Low	Low	Low
	Okamoto et al. [27]	two-level hierarchical clustering method	Topic extraction	High	Low	Low	Low
	Liu et al. [5]	clustering, modeling entities and weighted undirected bipartite graph	daily web news	Med	Low	Low	Low
	Tanev et al. [37]	automatic pattern learning	real-time news event extraction	High	Low	Low	Low
	Lei et al. [24]	SVM algorithm	Topic tracking	High	Low	Low	Low
Knowledge	Waterman et al. [39]			Low	Med	High	Med
	Beata [25]			Low	High	High	Med
	klaussner et al. [22]			Low	High	High	Med
	IJntema et al. [20]			Low	Med	High	Med
	Hearst et al. [16]	lexico-semantic patterns based on financial ontology	extracting financial events from RSS news feeds	Low	High	High	Med
	Jacobs et al. [21]			Low	Med	High	Med
	Hogenboom et al. [18]	use of lexico-semantic patterns	economic event extraction	Low	Med	High	Med
Hybrid	Piskorski et al. [29]	bootstrap a weakly supervised pattern learning algorithm with clusters	extract violence incidents from online news	N/A	Med	Med	Low
	Kenji et al. [32]	combine rule-based and data-driven NLP techniques	extraction of grammatical relations	Med	Med	High	Med
	Camiano et al. [9]	-statistical methods, rules based methods	resolve issues caused by the lack of expert knowledge	High	Med	Med	Med
	Pakhomov et al. [28]	statistical methods with lexical knowledge	reinforce statistical methods	Med	Med	Med	Med
	Lee et al. [23]	grammar-based statistical method	e-news summarization	Med	Med	Med	Med
	Chun et al. [15]	lexico-syntactic patterns, combined with term co-occurrences	Biomedical events	Med	Med	Med	Med

4 Conclusions

We present in this survey the main approaches in current literature, for event extraction from text. As shown, data-driven approaches (corpus based approaches) require a huge amount of data to discover statistical relations through the use of quantitative methods such as probabilistic modeling, information theory, and linear algebra to develop models that approximate linguistic phenomena, So these approaches require little domain knowledge and expertise. The main advantage of corpus based methods is that we don't need expert knowledge but we get low interpretability as a result. For the knowledge-driven approaches, we rely basically on patterns developed by experts but we need also a little amount of data to develop these patterns. Pattern based approaches gives better results with high interpretability but can't deal with huge amount of data when we are looking for the extraction of various types of events. The resort to hybrid approaches that combine knowledge-driven and data-driven approaches seems to be a great solution to remedy drawbacks of each family approach and get the advantages of both techniques: patterns based and corpus based methods.

Bibliography

[1] *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*, 2005.

[2] SPARQL query language for RDF, W3C recommendation, 2008.

[3] K. C. H. V. Aaltonen, S and A. Heinze. Social media in europe: Lessons from an online survey. Worcester College, Oxford, UK, 2013. 18th UKAIS Annual Conference: Social Information Systems.

[4] M. Adedoyin-Olowe, M. M. Gaber, and F. T. Stahl. A survey of data mining techniques for social media analysis. *CoRR*, abs/1312.4617, 2013.

[5] P. Anantharam, P. Barnaghi, K. Thirunarayan, and A. Sheth. Extracting city traffic events from social streams. *ACM Trans. Intell. Syst. Technol.*, 6(4):43:1–43:27, July 2015.

[6] T. Baldwin. Social media: Friend or foe of natural language processing? In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 58–59, Bali,Indonesia, November 2012. Faculty of Computer Science, Universitas Indonesia.

[7] J. Borsje, F. Hogenboom, and F. Frasincar. Semi-automatic financial events discovery based on lexico-semantic patterns. *Int. J. Web Eng. Technol.*, 6(2):115–140, 2010.

[8] Z. Chen, D. V. Kalashnikov, and S. Mehrotra. Exploiting context analysis for combining multiple entity resolution systems. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 207–218. ACM, 2009.

[9] P. Cimiano and S. Staab. Learning by googling. *SIGKDD Explor. Newsl.*, 6(2):24–33, 2004.

[10] H. Cunningham, D. Maynard, and V. Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, November 2000.

[11] G. De Pauw, G.-M. de Schryver, and P. Wagacha. Data-driven part-of-speech tagging of kiswahili. In *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, pages 197–204. Springer Berlin Heidelberg, 2006.

[12] M. Dredze and J. Wallenberg. Icelandic data driven part of speech tagging. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Short Papers*, pages 33–36, 2008.

[13] H. Farah. *Extraction de concepts et de relations entre concepts Ã partir des documents multilingues : Approche statistique et ontologique dissertation*. PhD thesis, Institut Nationale des Sciences AppliquÃ©es de Lyon, Lyon, France, 2009.

[14] B.-J. . L. L. Frasincar, F. A semantic web-based approach for building personalized news services. *International Journal of E-Business Research (IJEBR)*, 5:19, 2009. 3.

[15] R. Grishman. Information extraction: Capabilities and challenges. *Lecture Notes*, 2012.

[16] M. A. Hearst. Automated discovery of wordnet relations. *WordNet: an electronic lexical database*, pages 131–153, 1998.

[17] F. Hogenboom, F. Frasincar, U. Kaymak, and F. D. Jong. An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011). Volume 779 of CEUR Workshop Proceedings., CEURWS.org (2011)*, 2011.

[18] F. Hogenboom, A. Hogenboom, F. Frasincar, U. Kaymak, O. van der Meer, K. Schouten, and D. Vandic. Speed: A semantics-based pipeline for economic event detection. In J. Parsons, M. Saeki, P. Shoval, C. Woo, and Y. Wand, editors, *Conceptual Modeling ER 2010*, volume 6412 of *Lecture Notes in Computer Science*, pages 452–457. Springer Berlin Heidelberg, 2010.

[19] W. IJntema, J. Sangers, F. Hogenboom, and F. Frasincar. A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15(3), 2012.

[20] W. IJntema, J. Sangers, F. Hogenboom, and F. Frasincar. A lexico-semantic pattern language for learning ontology instances from text. *J. Web Sem.*, 15:37–50, 2012.

[21] P. S. Jacobs, G. R. Krupka, and L. F. Rau. Lexico-semantic pattern matching as a companion to parsing in text understanding. In *Proceedings of the Workshop on Speech and Natural Language*, pages 337–341, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics.

[22] C. Klaussner and D. Zhekova. Lexico-syntactic patterns for automatic ontology building. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 109–114, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee.

[23] C.-S. Lee, Y.-J. Chen, and Z.-W. Jian. Ontology-based fuzzy event extraction agent for chinese e-news summarization. *Expert Syst. Appl.*, 25(3):431–447, 2003.

[24] Z. Lei, L.-D. Wu, Y. Zhang, and Y.-C. Liu. A system for detecting and tracking internet news event. In Y.-S. Ho and H. J. Kim, editors, *PCM (1)*, volume 3767 of *Lecture Notes in Computer Science*, pages 754–764. Springer, 2005.

[25] B. Megyesi. *Data-Driven syntactic analysis methods and applications for Swedish*. PhD thesis, Doctoral dissertation Departement of Speech, Music and Hearing KTH, Kungliga Tekniska Hogskolan, 2002.

[26] C. S. Nicole B. Ellison and C. Lampe. The benefits of facebook friends: Social capital and college students use of online social network sites. *Computer Mediated Communication*, 12, July 2007.

[27] M. Okamoto and M. Kikuchi. Discovering volatile events in your neighborhood: Local-area topic extraction from blog entries. In G. G. Lee, D. Song, C.-Y. Lin, A. N. Aizawa, K. Kuriyama, M. Yoshioka, and T. Sakai, editors, *AIRS*, volume 5839 of *Lecture Notes in Computer Science*, pages 181–192. Springer, 2009.

[28] S. Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 160–167, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[29] J. Piskorski, H. Tanev, and P. O. Wennerberg. Extracting violent events from on-line news for ontology population. In W. Abramowicz, editor, *BIS*, volume 4439 of *Lecture Notes in Computer Science*, pages 287–300. Springer, 2007.

[30] V. Punyakanok, D. Roth, and W.-t. Yih. The importance of syntactic parsing and inference in semantic role labeling. *Comput. Linguist.*, 34(2):257–287, 2008.

[31] D. Rusu, J. Hodson, and A. Kimball. Unsupervised techniques for extracting and clustering complex events in news. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 26–34, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[32] K. Sagae, A. Lavie, and B. MacWhinney. Combining rule-based and data-driven techniques for grammatical relation extraction in spoken langugage. In *In Proceedings of the Eighth International Workshop in Parsing*, pages 153–162, 2003.

[33] A. Saval, M. Bouzid, and S. Brunessaux. A semantic extension for event modelisation. In *Tools with Artificial Intelligence, 2009. ICTAI '09. 21st International Conference on*, pages 139–146, Nov 2009.

[34] V. Soulignac. *Système informatique de capitalisation de connaissances et d'innovation pour la conception et le pilotage de systèmes de culture durables.* Theses, Université Blaise Pascal - Clermont-Ferrand II, Oct. 2012.

[35] S. Staab, M. Erdmann, and A. Maedche. Engineering Ontologies using Semantic Patterns. Seattle, 2001.

[36] S. Staab, M. Erdmann, and A. Maedche. Engineering ontologies using semantic patterns, 2001.

[37] H. Tanev, J. Piskorski, and M. Atkinson. Real-time news event extraction for global crisis monitoring. In *Proceedings of the 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems*, pages 207–218, Berlin, Heidelberg, 2008. Springer-Verlag.

[38] C. Walker, S. Strassel, J. Medero, and K. Maeda. Ace 2005 Multilingual Training Corpus. *Linguistic Data Consortium, Philadelphia*, 2006.

[39] S. A. Waterman. Structural methods for lexical/semantic patterns, 1993.

[40] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. J. Cunningham. Weka: Practical machine learning tools and techniques with java implementations, 1999.

[41] D. Zhou, L. Chen, and Y. He. A simple bayesian modelling approach to event extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705, Baltimore, Maryland, June 2014. Association for Computational Linguistics.