EVENT DETECTION AND CLASSIFICATION IN NATURAL TEXTS

Zoltán Subecz[1*]

[1] Department of Information Technology, GAMF Faculty of Engineering and Computer Science, John von Neumann University, Hungary

Keywords:
Artificial Intelligence
Machine Learning
Data Mining
Text Mining
Information extraction

Article history:
Received 08 September 2018
Revised 18 October 2018
Accepted 04 March 2019

Abstract

The detection and analysis of events in natural language texts plays an important role in several NLP applications such as summarization and question answering. In this study we introduce a machine learning-based approach that can detect and classify verbal and infinitival events in Hungarian texts. First we identify the multiword noun + verb and noun + infinitive expressions. Then the events are detected and the identified events are classified. For each problem, we applied binary classifiers based on rich feature sets. The models were expanded with rule-based methods too.

1. Introduction

Humans developed **natural language** to communicate; over past millennia, it has been the most efficient form of transferring the majority of information between individuals. With the advent of computing, large amounts of natural language text are stored in digital format. Computational linguistics helps link the significant power of the computer with the efficiency of communicating in natural language [5].

Natural language is an important communication tool and is widely used to disseminate knowledge and data. Natural languages are the languages that real people speak. Although language is patterned and organized, its processing is often complex and difficult. **Natural language processing (NLP)** is the computer processing of human language. It may span from speech to language understanding - from sounds to semantics.

As an essential part of **artificial intelligence (AI)**, natural language processing (NLP) investigates computationally effective algorithms capable of analyzing, understanding, and generating spoken, signed or written natural language [12, 1]. This field of computational linguistics concerned with developing methods for enabling computers to work with natural language, i.e. written texts or spoken language, the natural forms of human communication.

Information could be either ***structured*** or ***unstructured***. High voluminous amount of information is available in this world in the form of unstructured data which mostly exists in textual format. Unstructured data could exist in any form such as emails, literature papers, research papers, news articles, and blog posts. It can also exist in any human readable and spoken language.

Information extraction (IE) is an important task in the field of Natural Language Processing (NLP) that tries to extract information from semi-structured or un-structured machine readable documents and store it in a structured way that can be queried directly. The IE is usually considered as a subfield of Text Mining. IE has been applied to various applications such as question answering, information retrieval, conversational language understanding, machine translation and many more. Over the years, Information Extraction (IE) has become increasingly popular as a tool for a vast array of applications [4]. Some typical IE sub-tasks: named entity recognition, event extraction, coreference resolution, relation extraction.

[*] Corresponding author. Tel.: +36 76 516 411
E-mail address: subecz.zoltan@gamf.uni-neumann.hu

IE techniques have advanced from rule-based to statistical and machine learning based approaches. Rule-based methods use hand-coded patterns to extract information. While it is easy to implement and debug, they heavily rely on developers' heuristic and require lot of manual labor [3]. It usually has good precision but comparably low recall. Machine learning based approaches, on the other hand, are trainable, adaptable and extensible. With the development of human annotated corpora, machine learning based approaches have achieved significant progress.

Human languages refer not only to entities, but critically, also to situations. Therefore, various aspects of situations are worth analyzing in modeling linguistic meaning. The eventive dimension of information is fundamental for reasoning about how the world changes. The world is dynamic in its nature, and **events** are important aspects of everything that happens in this world. Things that happen and involve change (events), or situations that stay the same for a certain period of time (states) are related by their temporal reference.

Example for **events** and time in natural text:
*He **arrived** to the party at 8 p.m.*
*However, she had already **left**.*
*He **went** back home, after **talking** with some friends.*

Event extraction is an important task in Information Extraction (IE), which is a sub-field in Natural Language Processing (NLP) [2, 6, 14]. It has been applied to different genres (e.g., news articles, web blogs, tweets, etc.) and various applications (e.g., question answering, information retrieval, etc.). The goal of event extraction is to extract structure information for the events that are of interest from unstructured documents. It will be extremely valuable if we could automatically detect and extract such events effectively. In order to exploit this unstructured data, machine learning and text mining techniques can be used to recognize events.

2. Events

Time in language can be broken down into three primitives: times, **events** and temporal relations [11]. Viewing the temporal structure of a discourse as a graph, the times and events are the nodes and the relations the arcs.

According to the Cambridge English Dictionary, an event is "anything that happens, especially something important or unusual". In philosophy, events are objects in time or instantiations of properties in objects. However, a definite definition has not been reached, as multiple theories exist concerning events. The Oxford English Dictionary defines an event as "a thing that happens or takes place, especially one of importance".

Fiscus and Doddington [8] in the scope of the Topic Detection and Tracking project gave the following definitions event: event is “something that happens at some specific time and place along with all necessary preconditions and unavoidable consequences”. Becker et al. [2] adopt an event definition used in an earlier study on broadcast news: “An event is something that occurs in a certain place at a certain time”.

In this research, the description of events from TimeML (a temporal markup language) [13] is adopted, as follows: We consider "events" a cover term for situations that happen or occur. Events can be punctual or last for a period of time. We also con-sider as events those predicates describing states or circumstances in which something obtains or holds true.

There are often mentions of **negated events, conditional events** or **modal events**, which cannot be said to certainly "happen or take place" [13]. Further, events can be composed of many sub-events: for example, the Arab Spring lasted months and included multiple revolutions, each of which had a long history, a complex set of story threads all happening in parallel, a culmination and an aftermath. Events may be represented by a variety of lengths of expressions, ranging from document collections [15] to single tokens.

3. Event detection and classification

The detection and analysis of events in natural language texts plays an important role in several NLP applications such as summarization and question answering. In this paper we deal with the detection and classification of events that occur in natural language texts.

Though other parts of speech (e.g. noun, participle) can also denote events, the most events belong to verbs in texts; therefore we deal with verbal and infinitival events in this study. e.g. *A tanár **bement** a terembe.* (*The teacher **went** into the room*.) However not all verbs and infinitives can be considered as event-indicator (e.g. auxiliaries), thus special attention is needed to filter out them. *e.g. Haza **akarok** menni. (I **want** to go home.)*

The input of our system is a token-level labeled training corpus. The task was divided into three parts. First the single- and multiword verbal and infinitival expressions were picked out. Then from them the events were detected. Finally, the identified events were classified.

Our approach detects and classifies the events with machine learning techniques, which were expanded with rule-based methods. In our system we applied the Hungarian WordNet [10] for the semantic characterization of the examined words, and we disambiguated the polysemic inspected words with the Lesk algorithm [9]

4. The Corpus, the WordNet and Applied Software Packages

In our application we used one part of the Szeged Corpus [5], which contains 5,000 sentences from the following domains: business and financial news, fictions, legal texts, newspaper articles, compositions of pupils. From each of the five domains we selected the first 1,000 sentences.

Examples:

*A tanár **bement** a terembe.*	(event)	(The teacher **went** into the room)
*Haza **akarok** menni.*	(non event)	(I **want** to go home.)

The sentences were annotated by two annotators with the help of a linguist expert for the detection and classification. The inter-annotator agreement for detection was 87% and for classification it was 81% (simple percentage).

Wordnets are lexical databases in which words are organized into clusters based on their meanings, and they are linked to each other through different semantic and lexical relations, yielding a conceptual hierarchy (i.e. lexical ontology) of words. The Hungarian WordNet [10] comprises of over 40.000 synsets, out of which 2.000 synsets form part of a business domain specific ontology. The proportion of the different parts-of-speech in the general ontology follows that observed in the Hungarian National Corpus and includes approximately 19.400 noun, 3.400 verb, 4.100 adjective and 1.100 adverb synsets.

The J48 decision tree algorithm of the Weka* data mining suite was employed for machine learning. Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

For the linguistic processing of Hungarian texts the Magyarlanc [16] toolkit was used. The toolkit called magyarlanc aims at the basic linguistic processing of Hungarian texts. The modules of magyarlanc are: sentence splitter, tokenizer, POS tagger and lemmatizer, stopword filtering, dependency parser, constituency parser.

5. The Detection of Verbal and Infinitival Events

In this module we detected the verbal and infinitival events. Binary classification was performed for this task, which we expanded with rule based methods. For this module a separate classifier was created, where the event candidates were the verbs and infinitives.

The 5,000 sentences contain 10,628 verbs and infinitives, which were used as event candidates. The annotators labeled 6,479 of them as event.

5.1. Feature Set

The following features were defined for each event candidate

- **Surface features**: bigrams and trigrams: The character bigrams and trigrams of the beginning and end of the examined words. Besides them: word length, lemma length and the word position within the sentence.

* https://www.cs.waikato.ac.nz/ml/weka/

- **Lexical features**: binary feature: Is the examined word a copula or an auxiliary verb? Two lists were created with copulas and auxiliary verbs. These features indicate the presence of the lemma in these lists. Since the eventive nature of a word could be determined by the presence of a copula or an auxiliary verb before or after the word, these four binary features were used.
- **Morphological features**: Since the Hungarian language has rich morphology, therefore several morphology-based features were defined. We defined the MSD codes (morphological coding system) of the event candidates, using the next morphological features: type, mood(Mood), case(Cas), tense(Tense), person of possessor (PerP), number(Num), definiteness (Def). The following features were also defined: the verbal prefix, the examined word, the POS code and the POS codes of the previous and the subsequent words.
- **Syntactic features**: We defined the syntactic labels of the children of the examined event candidate (e.g. Subject, Object. . .)
- **Semantic features**: The Hungarian WordNet was used here, which contains 3,611 verbal synsets out of the all 42,292 synsets. The semantic relations of the WordNet hypernym hierarchy were used. *We applied the following method, which is new compared to the previous studies.* A separate model was created that without human interaction picked out synsets that are typically in the hypernym chains of events, or have an important role in the decision of the eventive nature. One of the advantages of our method is the automatic collection of the suitable synsets. Otherwise, finding all the required synsets with a simple method would be a complicated task because the events do not belong to some specific synsets in the diverse hypernym relation system of the WordNet. The second advantage of our method is that it can be applied generally, without modification, also to similar problems where it is necessary to find common hypernym intersections, relations for the group of given words in the WordNet hierarchy. It was applied also for the event classification. First we created a model, to which we collected the hypernyms of each event candidate as features during the training phase. On the basis of the features of the decision tree, the model picked out those synsets that are typically in the hypernym chains of events, or have an important role in the decision of the eventive nature. It picked out 95 synsets out of the 3,611 verbal synsets into a list. Then for the main model, these 95 binary features were added to the feature set. At the evaluation phrase we checked whether the event candidate belongs to the hyponyms of any of the collected synsets. Since several meanings can belong to a word form in the WordNet, therefore we performed word sense disambiguation (WSD) between the particular senses with the Lesk algorithm. [9]: Definition and illustrative sentences belong to the synsets in the WordNet. In the case of polysemic event candidates, we counted how many words from the syntactic environment of the event candidate can be found in the definition and illustrative sentences of the particular WordNet synset (neglecting stopwords). That sense was chosen which contained the highest number of common words.
- **Frequency features**: This feature group was applied as a new method as compared to previously published papers. As one of the features, we counted for each event candidate the rate of the cases when the particular word’s lemma is an event in the training set. As the second feature a similar rate was counted for the verbal prefix + lemma pair of each event candidate.

The number of features in each group: Surface: 7, Lexical: 6, Morphological: 10, Syntactic: 4, Semantic: 1–10, Frequency: 2

We completed our machine learning technique also with a **rule based method**. There were several expressions in the legal texts where the verb usually indicates event in other contexts, but not in the legal context. For example: *A törvény **kimondja**, hogy. . . (The law **states** that. . .)* We defined rules for such cases. An example for such a rule: If Subject = "law" And Candidate = "state" Then Candidate ≠ Event. We applied 68 such rules in the legal texts.

In the course of evaluation of event detection and classification, the precision, recall and F-measure metrics were used. We examined the significance of the particular feature groups too, then the model's performance on the five subcorpora separately.

Two baseline solutions were applied. At the first one, every verb and infinitive was treated as event. At the second one, only those verbs and infinitives were treated as event that is not copulas or auxiliary verbs.

6. The Classification of Verbal and Infinitival Events

After the detection of verbal and infinitival events we classified them. The classification was performed considering multiple aspects. First, we investigated the main verb types: actions, occurrences, existence and states. Out of them the action and occurrence categories are mostly related to events, therefore these two categories were focused on. **Examples** Action: *A postás **hoz** egy csomagot. (The postman **brings** a package.)* Occurrence: *A levél **leesett** a fáról. (The leaf has **fallen** from the tree.)* Within the 5,000 sentences, among the 6479 events there were 4,158 actions and 1,752 occurrences.

The actions and occurrences together constitute the main part of the events. We wanted to test our model, independently from the former classification, on smaller, but frequent categories. Hence for the second experiment two smaller categories were chosen: movement and communication. **Examples** Movement: *A gyerek **elment** az iskolába. (The child **went** to the school.)* Communication: *Tegnap telefonon **beszélgettünk**. (We **talked** on the phone yesterday.)* In the corpus there were 586 movement and 1,120 communication events.

The same feature set and feature selection methods were used as for the event detection.

Our machine learning technique was extended in the case of movements with a rule based method. Several expressions can be found that denote movement in most contexts, but in some cases they do not. For example: *Az árak szűk sávban **mozogtak**. (The prices **moved** in a narrow range.)* We defined rules for such cases. An example for such a rule: If Subject = "price" And Candidate = "move" Then Candidate ≠ Movement. We created baseline models for classifications too. We applied 11 such rules for movements.

7. Discussion, Conclusions

In this paper, we introduced our machine learning approach based upon a rich feature set, which can detect verbal and infinitival events in Hungarian texts and classify the identified events. We solved the problem in 3 steps. First, we identified the multiword noun + verb or noun + infinitive expressions. Then we detected the events, and classified the identified events. We tested our methods on 5 domains of the Szeged Corpus.

We applied for each problem binary classifiers based on rich feature sets. We expanded the models with rule based methods too. In this study we introduced new methods for this application area.

The detailed results will be published in a subsequent paper.

Acknowledgment

This publication is supported by EFOP-3.6.1-16-2016-00006 "The development and enhancement of the research potential at John von Neumann University" project. The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

References

[1] Allen, J. F. (1995). Natural language understanding (2nd ed.). Redwood City, CA, USA.: Benjamin-Cummings Publishing Co., Inc.,.

[2] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In WSDM'10.

[3] L. Chiticariu, Y. Li, and F. R. Reiss, \Rule-based information extraction is dead! long live rule-based information extraction systems!" in Proc. Conf. on Empirical Methods in Natural Language Process., Seattle, WA, 2013, pp. 827-832.

[4] Cowie, J., & Lehnert, W. (1996). Information Extraction. Communications of the ACM

[5] Csendes, D., Csirik, J.A., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Sojka, P., Kopecek, I., Pala, K. ˇ (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 41–47. Springer, Heidelberg (2004)

[6] F.P. Hogenboom, F. Frasincar, U. Kaymak, F.M.G. Jong: An overview of event extraction from text. Proceedings of Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), pp. 48-57. Aachen 2011

[7] Leon R.A. Derczynski: Determining the Types of Temporal Relations in Discourse, University of Sheffield, 2013

[8] Fiscus, J.G., Doddington, G.R.: Topic detection and tracking evaluation overview. Topic detection and tracking pp. 17–31 (2002)

[9] Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing. In: Computational Linguistics, and Speech Recognition. PrenticeHall, Upper Saddle River (2000)

[10] Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.) Proceedings of the Fourth Global WordNet Conference, GWC 2008, pp. 311–320. University of Szeged, Szeged (2008)

[11] Moens, M. and M. Steedman (1988), “Temporal ontology and temporal reference.” Computational linguistics, 14, 15–28.

[12] Moreno, L., Palomar, M., Molina, A., & Ferrandez, A. (1999). Introduccion al Procesamiento del Lenguaje Natural. Servicio de Publicaciones de la Universidad de Alicante.

[13] Pustejovsky, J.: The syntax of event structure. Cognition 41(1-3), 47 (1991)

[14] Pustejovsky, J., B. Ingria, R. Sauri, J. Castano, J. Littman, and R. Gaizauskas (2004), “The Specification Language TimeML.” In The Language of Time: A Reader, 545–557, Oxford University Press.

[15] Ritter, A., Etzioni, O., Clark, S., et al.: Open domain event extraction from Twitter. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1104–1112. ACM (2012)

[16] Zsibrita, János; Vincze, Veronika; Farkas, Richárd 2013: magyarlanc: A Toolkit for Morphological and Dependency Parsing