Event Detection and Classification in Hungarian Natural Texts

Zoltan Subecz
Department of Information Technology, GAMF Faculty of Engineering and Computer Science, John von Neumann University, Hungary

Doi:10.19044/esj.2019.v15n21p411 URL:http://dx.doi.org/10.19044/esj.2019.v15n21p411

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. This paper focuses on introducing a machine learning-based approach that can detect and classify verbal and infinitival events in Hungarian texts. First, the multiword noun + verb and noun + infinitive expressions were identified. Then the events are detected and the identified events are classified. For each problem, binary classifiers were applied based on rich feature sets. The models were expanded with rule-based methods.

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

Introduction

Humans developed **natural language** to communicate. Over the 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 (Derczynski, 2013).

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 (Moreno, 1999; Allen, 1995). This field of computational linguistics is

concerned with developing methods for enabling computers to work with natural language, i.e., written texts or spoken language, which is the natural forms of human communication.

Information can 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. This information is then stored 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, and it has been used as a tool for a vast array of applications (Cowie, 1996). Some typical IE sub-tasks include; entity recognition, event extraction, coreference resolution, and relation extraction.

Additionally, 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 a lot of manual labor (Chiticariu, 2013). 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** at 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) (Becker, 2010; Hogenboom, 2011; Pustejovsky, 2004). 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 structured information for the events that are of interest from unstructured documents. It will be extremely valuable if such events could be automatically detected and extracted effectively. In order to exploit this unstructured data, machine learning and text mining techniques can be used to recognize events.

Events

Time in language can be broken down into three primitives: times, **events,** and temporal relations (Moens, 1988). Viewing the temporal structure of a discourse as a graph, the times and events are the nodes and the relations is 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 (Fiscus, 2002) in the scope of the Topic Detection and Tracking project gave the following definitions of event: event is “something that happens at some specific time and place along with all necessary preconditions and unavoidable consequences.” Becker et al. (2010) adopts 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) (Pustejovsky, 1991) is adopted as follows: The "events" are considered as a cover term for situations that happen or occur. Events can be punctual or last for a period of time. The events are also considered as those predicates describing states or circumstances in which something obtains or holds true.

Negated events, conditional events or modal events are often mentioned, and it cannot be said to certainly "happen or take place" (Pustejovsky, 1991). 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 to single tokens (Ritter, 2012).

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. This paper will 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, this study will deal with verbal and infinitival events 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 them out, e.g., *Haza* ***akarok*** *menni (I* ***want*** *to go home.)*

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

The demonstrated approach detects and classifies the events with machine learning techniques, which were expanded with rule-based methods. In this system, the Hungarian WordNet (Miháltz, 2008) was applied for the semantic characterization of the examined words, and the polysemic inspected words were disambiguated with the Lesk algorithm (Jurafsky, 2000).

The Corpus, the WordNet, and Applied Software Packages

In the demonstrated application, one part of the Szeged Corpus (Csendes, 2004) was used which contains 5,000 sentences from the following domains: business and financial news, fictions, legal texts, newspaper articles, and compositions of pupils. Furthermore, the first 1,000 sentences were selected from each of the five domains.

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).

*WordNet*s 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 (Miháltz, 2008) 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 which is 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. Also, it contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

For the linguistic processing of Hungarian texts, the Magyarlanc (Zsibrita, 2013) 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, and constituency parser.

The Detection of Verbal and Infinitival Events

In this module, the verbal and infinitival events were detected. 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 as event.

Feature Set

The following features were defined for each event candidate

- **Surface Features (Bigrams and Trigrams):** The character reveals the bigrams and trigrams of the beginning and end of the examined words. It also shows the word length, lemma length, and the word position within the sentence.
- **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, several morphology-based features were defined. First, the next words were defined: stem and prefix+stem features. Then, the MSD codes (morphological coding system) of the event candidates were processed using the next morphological features: type, mood (Mood), case (Cas), tense (Tense), person of possessor (PerP), number (Num), and 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**: The syntactic labels of the children of the examined event candidate (e.g., Subject, Object) were defined.
- **Semantic Features**: The Hungarian WordNet was used here, which contains 3,611 verbal synsets out of all the 42,292 synsets. The semantic relations of the WordNet hypernym hierarchy were used. *The following method was applied, which is new compared to the previous studies*. A separate model was created without human interaction that 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 the demonstrated method is the automatic collection of the suitable synsets. Otherwise, finding all the required synsets with a simple method would be a complicated task. This is because the events do not belong to some specific synsets in the diverse hypernym relation system of the WordNet. The second advantage of this method is that it can be applied generally, without modification, to similar problems where it is necessary to find common hypernym intersections, which is associated to the group of given words in the WordNet hierarchy. It was applied also for the *event classification*. First, a model was created from which the hypernyms of each event candidate were collected 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, that was 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, word sense disambiguation (WSD) was performed between the particular senses with the Lesk algorithm (Jurafsky, 2000). Definition and illustrative sentences belong to the synsets in the WordNet. In the case of polysemic event candidates, it was counted that 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). The sentences which contained the highest number of common words were chosen.

Furthermore, the number of features in each group includes the following: Surface: 7, Lexical: 6, Morphological: 12, Syntactic: 4, Semantic: 1–10.

The demonstrated machine learning technique was also completed with **rule based methods**. 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. . .).* Rules were defined for such cases. An example for such a rule: If Subject = "law" And Candidate = "state" Then Candidate ≠ Event. 68 such rules were applied in the legal texts.

In the course of evaluation of event detection and classification, the precision, recall and F-measure metrics were used. The significance of the particular feature groups was examined too. The model's performance was also observed 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 verbs and infinitives that were neither copulas nor auxiliary verbs were treated as event.

Results: Event Detection

The following experiments on event detection were performed with 10 fold cross validation.

The first **baseline method** achieved an F-measure of 79.45, while the second one reached an F-measure of 84.37.

With only the WordNet feature used independently, the model achieved an F-measure of 91.84.

With the whole feature set, the model achieved the following scores: precision: **94.76, recall: 96.20, and F-measure: 95.48**.

The efficiency of the particular feature groups was examined with an **ablation analysis**. In this case, the particular feature groups were left out from the whole feature set and the model was trained on the basis of the residual features. The results can be found in Table 1. According to the results, the Morphological and Semantic features proved to be the most useful ones. The best result was achieved without the Surface features. Therefore, our further experiments were performed without them.

Table 1. Results of the ablation analysis - Event detection

Left out features	**Precision**	**Recall**	**F-measure**	**Difference**
Surface	94.52	96.50	95.50	+0.02
Lexical	94.67	96.16	95.41	−0.07
Morphological	94.74	96.17	95.45	**−1.05**
Syntactic	94.80	95.99	95.39	−0.09
Semantic	94.63	96.06	95.34	**−0.14**

Then the model was tested on verbs and without the rule based method. It resulted an F-measure of 94.75 with focus only on verbs. It resulted an F-

measure of 95.20 without the rule based method. Henceforward, the rule based method was used together with focus on verbs and infinitives.

The model's performance was also examined on each subcorpus. These results can be seen in Table 2. The model achieved the best performance on the *Business news* domain, and the lowest performance on the Legal corpus.

Table 2. Performance on the subcorpora – Event detection

Corpus	**Precision**	**Recall**	**F-measure**
Compositions	96.08	98.00	97.03
Legal	89.74	86.42	88.05
Fictions	95.45	97.35	96.39
Business news	97.86	98.56	**98.21**
Newspaper articles	96.71	97.35	97.03

The Classification of Verbal and Infinitival Events

After the detection of verbal and infinitival events, these were *classified*. The *classification* was performed considering multiple aspects. First, the main verb types were investigated: actions, occurrences, existence, and states. Out of all of them, the action and occurrence categories are mostly related to events. Therefore, these two categories were focused on. **Example** of **Action category**: *A postás **hoz** egy csomagot (The postman **brings** a package).* **Example of Occurrence category**: *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. Independently from the former classification, for the second experiment, the model was tested on the next smaller, but frequent categories: movement and communication. **Example of Movement category:** *A gyerek **elment** az iskolába (The child **went** to the school).* **Example of Communication category:** *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 for the event detection.

The demonstrated machine learning technique was extended in the case of movements with **rule based methods**. 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).* Rules were defined for such cases. An example for such rule: If Subject = "price" And Candidate = "move" Then Candidate ≠ Movement. Baseline models were created for classifications too. 11 such rules were applied for movements.

Results: Event Classification

The following experiments on event classification were performed with 10 fold cross validation.

In the action-occurrence classification task, the *baseline* model treated all events as action. The model achieved an F-measure of 78.38. In the movement and communication classification task, for the *baseline* model, 11 frequent verbs were selected that denote movement and 16 frequent verbs that denote communication events. The model treated only these events belonging to the particular category. The model achieved an F-measure of 49.15 for movement and 45.07 for communication.

Henceforward, the following abbreviations indicate the given categories:
A: Action, O: Occurrence, M: Movement, C: Communication
With only the WordNet feature used independently, the model achieved F-measures of A: 86.63; O: 66.00; M: 65.64; C: 81.24.
With the whole training set, the model achieved F-measures of A: 87.06; O: 73.43; M: 68.51; C: 81.57

Also, the significance of the particular feature groups was examined with an ablation analysis. In this case, the particular feature groups were left out from the whole feature set and the model was trained on the basis of the residual features. The results can be found in Table 3. According to the results, the Morphological and Semantic features proved to be the most useful ones.

Table 3. The results of ablation - F-measure - Event classification

Left out features	**Action**	**Occur-rence**	**Move-ment**	**Commu-nication**	**Difference**
Surface	87.02	73.58	68.40	81.13	−0.04/+0.15/−0.11/−0.44
Lexical	86.90	73.09	68.37	80.32	−0.16/−0.34/−0.14/−1.25
Morpho-logical	**84,65**	**70,58**	**59,54**	**78,91**	**-1,50/-2,35/-7,83/-1,72**
Syntactic	85.58	73.54	68.54	80.74	−1.48/+0.11/+0.03/−0.83
Semantic	**86.21**	**72.52**	**66.02**	**80.22**	**−0.85/−0.91/−2.49/−1.35**

The model's performance was examined on each sub-corpus. These results can be seen in Table 4. According to the average results, the model achieved the best performance on the Business news domain, and the lowest performance on the Newspaper articles corpus.

Table 4. *Performance on the sub-corpora - Event classification (F-measure)*

Corpus	Action	Occurrence	Movement	Communication
Compositions	85.32	56.67	86.96	75.68
Legal	84.40	71.43	66.67	84.85
Fictions	85.71	60.32	70.27	72.34
Business news	**88.89**	**92.86**	**62.37**	**85.71**
Newspaper articles	83.09	47.76	58.22	70.18

Additional Experiments for Event Classification

In the next two paragraphs, the improved results are marked bold compared to the outcome of the ablation analysis.

The feature set was extended with bag-of-words features. First, the lemmas of the syntactic dependents of the particular event candidate were used as bag-of-words. The extended model achieved F-measures of **A: 87.18; O: 74.01; M: 69.20;** C: 81.61 with 10 fold cross validation.

Then similar to the previous case, the lemmas of the syntactic dependents of the particular event candidate together with the relationship type were used as bag-of-words. For example: SUBJ-teacher. This extended model achieved F-measures of **A: 87.63; O: 74.04; M: 68.92;** C: 81.69 with 10 fold cross validation.

Conclusion

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

For each problem, binary classifiers were applied, based on rich feature sets. The models were expanded with rule based methods too. In this study, new methods were introduced for this application area. According to our best knowledge, this is the first result for detection and classification of verbal and infinitival events in Hungarian natural language texts. The model's feature set was tested with an ablation analysis, and the model's performance on 5 sub-corpora. Evaluating them on test databases, the demonstrated algorithms achieved competitive results as compared to the current English results. An F-measure of 95.5 was achieved for detection, and F-measure of 87.63; 74.04; 69.20 and 82.34 was achieved for the four classifications.

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. Becker, H., Naaman, M., & Gravano, L. (2010). *Learning similarity metrics for event identification in social media*. In WSDM'10.
3. http://web2.cs.columbia.edu/~gravano/Papers/2010/wsdm10.pdf
4. Chiticariu, L., Li, Y., & Reiss, F.R. (2013). *Rule-based information extraction systems* in Proc. Conf. on Empirical Methods in Natural Language Process., Seattle, WA, pp. 827-832.
5. http://www.aclweb.org/anthology/D13-1079
6. Cowie, J. & Lehnert, W. (1996). *Information Extraction*. Communications of the ACM.
7. Csendes, D., Csirik, J.A., & Gyimóthy, T. (2014). *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.
8. Fiscus, J.G. & Doddington, G.R. (2002). *Topic detection and tracking evaluation overview.* Topic detection and tracking pp. 17–31
9. https://pdfs.semanticscholar.org/f753/eaae780e5731d29ef4fbce02e58584c39792.pdf
10. Hogenboom, F.P., Frasincar, F., & Kaymak, U. (2011). 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.
11. Jurafsky, D. & Martin, J.H. (2000). Speech and Language Processing: *An Introduction to Natural Language Processing*. In: Computational Linguistics, and Speech Recognition. PrenticeHall, Upper Saddle River.
12. Leon, R.A.(2013). Derczynski: *Determining the Types of Temporal Relations in Discourse*, University of Sheffield,http://etheses.whiterose.ac.uk/4068/1/phdthesis.pdf
13. Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., & Váradi, T. (2008). *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.

14. Moens, M. & Steedman, M. (1988). *Temporal ontology and temporal reference*. Computational linguistics. 14, 15–28.
15. https://aclanthology.info/pdf/J/J88/J88-2003.pdf
16. Moreno, L., Palomar, M., Molina, A., & Ferrandez, A. (1999). *Introduccion al Procesamiento del Lenguaje Natural*. Servicio de Publicaciones de la Universidad de Alicante.
17. Pustejovsky, J. (1991). *The syntax of event structure*. Cognition 41(1-3), 47 (1991)
18. http://jamespusto.com/wp-content/uploads/2018/08/Gaiz-02-1.pdf
19. Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., & Gaizauskas, R. (2004). *The Specification Language TimeML*. In The Language of Time: A Reader, 545–557, Oxford University Press.
20. Ritter, A., Etzioni, O., Clark, S. (2012). *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
21. http://homes.cs.washington.edu/~mausam/papers/kdd12.pdf
22. Zsibrita János, Vincze Veronika, & Farkas Richárd, (2013). *magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian*. In: Proceedings of RANLP 2013, pp. 763-771.
23. http://publicatio.bibl.u-szeged.hu/3981/1/Zsibrita-Vincze-Farkas.pdf