Event Detection and Classification, and Semantic Role Labeling

PhD Thesis Summary

Zoltán Subecz

2019

Supervisor: Prof. Dr. János Csirik

University of Szeged
Doctoral School of Computer Science

1. Introduction

Our present time is often referred to as *Information Age*. The amount of information that has become available via the Internet is growing exponentially, for example, in the form of news, reports and messages. Accessing, searching information and improving and automating information processing are new challenges to gain more advantage from this valuable knowledgebase.

With the advent of computers huge amounts of the natural texts are stored in digital form. *Natural Language Processing* (NLP) is the processing of human languages by means of computers (Jurafsky & Martin, 2009) ranging from speech processing to semantics. *Information extraction* (IE) is an important part of NLP. It collects information from unstructured or semi-structured documents and stores them in a structured form. *Event Extraction* (EE) is an important subtask of IE. Its goal is to extract event information from unstructured documents.

Event information has become overdue for many NLP applications, such as Question Answering (Moldovan, Clark, & Harabagiu, 2005), Summarization (Mani & Shiffman, 2005), Information Retrieval (Alonso, Gertz, & Baeza-Yates, 2007) and Information Extraction (Surdeanu, Harabagiu, Williams, & Aarseth, 2003). Question answering research has showed (Sauri, Knippen, Verhagen, & Pustejovsky, 2005) that most web searches are event-related. *EE* is used in several areas in everyday life, such as politics, finance, economy, commerce, market research, decision support and healthcare. Parliamentary elections, announcements, management changes and acquisitions imply events. Events generate trading signals in the stock markets because financial markets are very sensitive to important news.

The examples of the dissertation were taken (where it was possible) from the Szeged Corpus (Csendes, Csirik, & Gyimothy, 2004).

An example of textual event sequences:

*We **got** there in the end because one of my classmate's father **came** and **took** us in his car. (De végül is **odaértünk**, mert **jött** az egyik osztálytársam apukája kocsival és **elvitt** minket.)*

From these sentences the reader can recreate the following reality: There is an event (came), which took place at a given time. And there is another event (took), which happened after the first event. The first event mentioned (got there) took place after the previous two events. These events mean status changes for those involved in the story. It would be very valuable if events could be detected and extracted efficiently and automatically.

The task of event detection is the identification of event-occurrences in texts. Every expression is treated as an event-occurrence that indicates an event or state which can be linked to a given time or interval. The detection, analysis and temporal relationship of events in texts are important factors in getting to know the textual content.

Most events belong to verbs in texts and verbs usually denote events. For example: *We **went** down for a swim in the evening as we **had agreed*** (*Este, ahogy **megbeszéltük**, **lementünk** fürödni*). Nevertheless, not all verbs and infinitives can be treated as event indicators (for example: is, was, will be, remain, auxiliaries), thus their filtering requires special attention. For example: *We **wanted** to go there this year, but the holiday house had been sold, so it didn't work out.* (*Ebben az évben is oda **akartunk** menni, de a nyaralót eladták, így nem jött össze.*)

But other parts of speech (e.g. noun, participle) can also denote events. For example: ***Running** was followed by **gymnastics** for 15 minutes, which we still did together* (*A **futás** után következett a **torna** negyed órán át, amit még közösen csináltunk*). Nominal events have two parts: deverbal and non-deverbal events. Deverbal events: *running (futás)*, *writing (írás)*. Non-deverbal nouns: *war (háború), feast (ünnep)*.

Deverbal nouns have two main types: events and results. These nouns are often ambiguous. Some of them are events in some sentences; others are results in other sentences. For example, the noun *writing* is an event in the following sentence:

However, time passed very quickly both while waiting and ***writing*** *the entrance exam.* (*Azonban az idő hamar elszaladt, a várakozás és a felvételi* ***írás****a közben egyaránt.*)

However, it isn't an event but a result in the following sentence: *Then, we visited the museum under the fortress, where several kinds of weapons, fighting tools,* ***writings*** *could be seen.* (*Ezután megnéztük a vár alatt lévő múzeumot, ahol külön-féle fegyvereket, harci eszközöket,* ***írás****okat lehetett látni*). Because of the ambiguity, words analysis is insufficient; also, the context must be analyzed.

Besides event detection another important task is to determine the roles of the events discovered. It is known as Semantic Role Labelling (SRL). It is the task of natural language processing to detect the semantic arguments of a sentence predicate and to classify them according to specific roles. *Semantic roles* are logical relationships between events and participants.

NLP involves the discovery of the text structure on morphological, syntactic and semantic levels. Besides syntactic analysis it is also important to reveal semantic relations (Carreras, 2005). Semantic information describes the relationship between the predicate and the syntactic components. The identification of these relationships is important in answering questions such as Who?, What?, Where?.

For example: *I noticed an old man reading a newspaper and* ***eating*** *a crescent in his car.* (*Észre vettem egy bácsit, aki éppen újságot olvasott és kiflit* ***evett*** *az autójában*.) The verb eat has three roles in this case. **The agent of eating** = old man, **the thing eaten** = crescent, **the eating place** = in his car.

Semantic role labelling is halfway between syntax and semantics. It is more a semantic task than part of speech tagging or syntactic analysis, but less semantic than information extraction or question answering. Related works (Christensen,

Mausam, & Etzioni, 2010) have showed that we could improve the efficacy of several higher level tasks making use of the results of an SRL system.

This dissertation is concerned with computer processing of events expressed in natural languages. Its main tasks are *event detection, event classification and the labelling of their semantic roles.*

2. The achievements of the thesis

The main achievements of this dissertation are summarized below along with the related publications.

I considered my **main task** in all three research areas to develop in detail feature groups that **take into account the *characteristics of the Hungarian language***. These were the ***morphological* and the *dependency-tree based feature groups*.** Since the Hungarian language is a morphologically rich language, I paid great attention to the ***morphological feature group***. And since the Hungarian language has free word order and dependency-tree based representations are well-suited for the analysis of languages with free word order, I paid particular attention to the ***dependency-tree based feature group*** as well. These feature groups significantly improved the results for the Hungarian texts.

In all three themes I used the following sources and methods:

- In my applications one part of the Hungarian Dependency Treebank (Vincze, Szauter, Almási, Móra, Alexin, & Csirik, 2010) was used from the following areas: business and financial news, fictions, legal texts, newspaper articles, compositions of pupils. I tested the model's performance on each subcorpus.
- In my systems I used the Hungarian WordNet (Miháltz at al. 2008) for the semantic characterization of the words examined, where the semantic relations of the WordNet hypernym hierarchy were used. Since a word form may have more than one sense in the WordNet, I performed word sense disambiguation

(WSD) of the particular senses using the Lesk algorithm (Jurafsky & Martin, 2009).

- I examined the efficacy of the particular feature groups using ablation analysis.
- For syntactic characterization the dependency-tree based representation was used. In this case not only the directly connected words were examined for the candidate verb, but also the relation between the candidate and the words farther away in the tree was analyzed (theme 2 and 3).

The key role the **morphological** and the **dependency-tree based syntactic** feature groups have in more cases confirms that besides the features used for English texts, it is useful to define **features for Hungarian** text analysis which take into account the **characteristics of the Hungarian language**.

2.1. The detection and classification of verbal and infinitival events in natural language texts.

Most events belong to verbs in texts and verbs usually denote events. That is why I dealt with the detection and classification of verbal and infinitival events separately (Subecz Z. , 2014). I introduced my rich feature set based machine-learning approach that can automatically detect and classify verbal and infinitival events.

Most works deal only with certain events (e.g. business) or more specific ones (e.g. acquisitions). I dealt with all types of verbal and infinitival event detection and classification.

I divided the task into three parts. First, the multiword noun + verb and noun + infinitive expressions were identified, then, the events were detected from them. Afterwards, the events found were classified. My approach detects and classifies events with machine learning techniques and has been expanded with a rule based method on the Legal Corpus.

I used a rich feature set based classifier in my model with the following feature groups: surface, lexical, morphological, syntactic (dependency-tree based representation) and semantic (WordNet) features.

A new model has been created for the WordNet feature group, which picks out the synsets that are typically found in the hypernym chains of events, then, the synsets picked out have been used in the main classifier. Also, in the WordNet feature group I tested my model with and without the Lesk algorithm.

For morphological analysis I also used the RFSA morphological parser of the magyarlanc linguistic toolkit (Zsibrita, Vincze, & Farkas, 2013).

For the morphological and syntactical (dependency-tree based representation) features I applied the bag of words model for the characterization of word groups.

I tested the model's performance separately for verbs and infinitives.

Besides the main examinations cross-domain measurements were carried out. In this case, the model trained on the source corpus was evaluated on the target corpus. I represented the similarity between domains in graph.

I examined using measurements how changes in corpus size modifies results.

After the detection of verbal and infinitival events I *classified* them. The classification was performed according to several criteria. First, I investigated the main verb types: actions, occurrences, existence and states. Out of them the *action* and *occurrence* categories are mostly related to events, therefore I focused on these two categories. I tested my model on smaller, but frequent categories: movement and communication.

The following points have been verified regarding the detection and classification of verbal and infinitival events (Thesis 1):

- *I have showed that in this area the best performing feature groups are the morphological, dependency-tree based syntactic and semantic groups.*
- *I have justified that applying the rule based method to the Legal Corpus improves the results of the machine learning system.*

- *I have showed that applying the Lesk algorithm to the WordNet feature group improves the results.*
- *I have proved that applying the bag of words model at the morphological and syntactic (dependency-tree based representation) features improves the results for the following word groups: word stem, prefixes and suffixes; relations and relation-lemmas at the parse tree.*
- *I have showed that regarding detection the model performs better for verbs than for infinitives.*
- *I have justified that regarding detection and classification Compositions of pupils, Fictions, Financial news and Newspaper articles were very similar to each other; legal texts were significantly different.*
- *I have proved that regarding detection and classification, increasing the corpus size improves the results, but the added value constantly declines.*

I presented the results of this Thesis in the following publications: (Subecz Z. , 2014), (Subecz & Csák, 2014). The co-author of the last publication provided the linguistic background.

2.2. Automatic detection of nominal events in Hungarian texts with dependency-tree and constituency-tree based representations and the WordNet

Besides verbs other parts of speech (e.g. nouns, participles) can also denote events. Among them nominal events are the most frequent, therefore I dealt with nominal event detection in detail (Subecz Z. , 2016). I introduced my machine learning approach based upon a rich feature set, which can detect nominal events in Hungarian texts using dependency-tree and constituency-tree based representations and the WordNet.

I used a classifier based upon a rich feature set with the following feature groups: Surface, Morphological, Dependency-tree based, Constituency-tree based, Semantic (WordNet), Bag of words, List and Combined features. For nominal event detection I also implemented a Named Entity Recognition System (Szarvas, Farkas, & Kocsor, 2006).

For syntactic characterization I used dependency-tree based and constituency-tree based representations and compared their efficacy.

The performance of the model was tested for deverbal and non-deverbal nominal events too.

The bag of words model was used for the characterization of word groups for these features: tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy; the words around the candidate in the sentence.

Two morphological parsing were used for morphological analysis. Two data mining algorithms were implemented and compared (Decision Tree and Support Vector Machine (SVM) algorithms).

I compared the results by forming groups from the candidates and without groups.

Besides the main features the following additional methods were used, which improved the results: forming candidate groups; feature weighting.

The following points have been verified for nominal event detection (Thesis 2):

- *I have showed that in this area the best performing feature groups are the semantic and the bag of words groups.*
- *I have proved that the bag of words model can be applied effectively in the case of the following word groups: tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy; the words around the candidate in the sentence.*
- *I have showed that, in this area, the dependency-tree based representation produces better results than the constituency-tree based one.*

- *I have justified that nominal event detection produces better results by forming groups from the candidates than by handling all candidates in one group only.*

I presented the results of this Thesis in the following publications: (Subecz Z. , 2016), (Subecz Z. , 2017a), (Subecz Z. , 2017b).

2.3 Automatic semantic role labelling of events

Besides event detection, the *labelling of the semantic relations* of events is an important task (Semantic Role Labelling, SRL). The detection of the events and their semantic roles can be utilized in several areas of natural language processing, for example, in summarization, machine translation and question answering.

I introduced my machine learning approach based upon a rich feature set, which can automatically label semantic roles of events (Subecz Z. , 2015a). I searched for roles for target words of verbal and infinitival events.

In semantic role labelling I dealt with the frames of *company purchases* and *stock market news*. In both cases, several domain-specific roles were labelled (5 and 8 roles in the particular frames).

In my model I used a classifier based upon a rich feature set with the following feature groups: Surface, Morphological, Syntactic (dependency-tree based representation) and Semantic (WordNet) features.

In the WordNet feature group I tested my model with and without the *Lesk algorithm.*

The *bag of words model* was used for the morphological, syntactic and semantic (WordNet) feature groups for the following word groups: word stem, prefixes and suffixes; tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy;

In the *company purchases* domain I examined my model's performance sorting out the targets into two groups: customer-centric and seller-centric groups.

I *left out the infrequent feature-entities* from the classifier, thus the vectorspace-size was reduced. I have investigated the effect of this leaving out.

In simpler cases the roles have direct syntactic relationship with the target word. I examined the model's performance if it dealt only with these directly connected candidates.

On each domain I have **searched for the roles** that the model can detect most **efficiently**.

The following points have been verified for the semantic role labelling of events (Thesis 3):

- *I have showed that for the target words of verbal events we can effectively search roles with machine learning techniques.*
- *I have justified that in this area, the best performing feature groups are the syntactic and the morphological groups. These groups have had positive impact on all examined roles.*
- *I have showed that in addition to the above, using semantic features also improves the results therefore it is recommended that the WordNet be applied here too.*
- *I have proved that applying the Lesk algorithm to the WordNet feature group improves the results.*
- *I have justified that applying the bag of words model to the morphological, syntactic and semantic features improves the results for the following word groups: word stem, prefixes and suffixes; tokens of subtrees; tokens between two nodes in the parse-tree; the synsets between two nodes in the WordNet hypernym hierarchy.*
- *I have showed that the model achieves better results if it deals with farther candidates in the dependency tree as well.*

- *I have proved that leaving out the infrequent feature-entities from the classifier improves the results.*
- *I have showed that on the purchases of companies domain the Price (Ár) and the Item (Árú) and on the e news from stock markets domain the Price (Ár) and the Shift-direction (Elmozdulás-irány) roles can be detect with most efficiency.*

I presented the results of this Thesis in the following publications: (Subecz Z. , 2015a), (Subecz Z. , 2015b)

The relationship between the publications and the described thesis statements is demonstrated in the following chart:

			Thesis		
			1	**2**	**3**
MSZNY	2014	Subecz Z. et al., 2014a	•		
TSD	**2014**	Subecz Z., 2014b	•		
MSZNY	2017	Subecz Z., 2017		•	
ICIST	**2016**	Subecz Z., 2016		•	
Informatics	**2017**	Subecz Z., 2017b	•	•	
MSZNY	2015	Subecz Z., 2015a			•
TSD	**2015**	Subecz Z., 2015b			•

2.5. Conclusions and Future Work

In this thesis I focused on the detection and classification of events and the labelling of their semantic roles. I aimed to **develop rich feature groups**, where many kinds of features were tested and compared. I **considered my main task** to develop in detail feature groups that take into account the **characteristics of the Hungarian language**. Based on the main contributions, I can argue that:

- for event detection and classification and semantic role labelling I successfully applied supervised machine learning approaches;

- the important parts of events in texts can be extracted automatically with machine learning methods;
- besides the features that were used for English texts it is useful to develop features for the analysis of Hungarian texts that take into account the characteristics of the Hungarian language; these are the morphological and the dependency-tree based feature groups;
- besides the morphological and syntactic features it is useful to apply semantic features for event detection and classification, and semantic role labelling;
- the application of the bag of words model is useful for the characterization of word groups, for example in the following cases: tokens of subtrees; tokens between two nodes in the parse-tree; synsets between two nodes in the WordNet hypernym hierarchy; the words around the candidate in the sentence;
- the convenient grouping and pre-processing of the candidates improves classification results;
- applying the Lesk algorithm to the WordNet feature group improves the results;
- in several cases, applying the rule based method improves the results of the machine learning system;
- when extracting events information, increasing the corpus size improves the results but the added value constantly declines;

Events information extraction has become more and more timely for many NLP applications, such as Question Answering, Summarization, Information Retrieval and Information Extraction. Question answering research has showed that most web searches are event-related. Summarization also requires event information and makes use of the relative sequence of events.

In the future, I would like to improve my systems by conducting a detailed analysis of the effects of the features included and developing systems for other languages as well by adapting language specific features. Later, I would like to generalize the features to achieve a language-independent event extraction system. Moreover, I

plan to integrate my event detection and classification, and semantic role labelling applications into one complex system.

I believe that my research on automatic event detection and classification, and semantic role labelling can be successfully exploited in several NLP tasks and it will contribute to developing novel approaches in many areas of natural language processing.

References

Alonso, O., Gertz, M., & Baeza-Yates, R. (2007). On the Value of Temporal Information in Information Retrieval. *ACM SIGIR Forum, volume 41* (pp. 35-41). New York, NY, USA: ACM.

Carreras, X. (2005). *Learning and Inference in Phrase Recognition, Doctoral thesis.* Catalunya: Universitat Politècnica de Catalunya (UPC).

Christensen, J., Mausam, S., & Etzioni, O. (2010). Semantic role labeling for open information extraction. *Proceeding FAM-LbR '10 Proceedings of the NAACL HLT 2010 First International* (pp. 52-60). Stroudsburg, PA, USA: Association for Computational Linguistics.

Csendes, D., Csirik, J., & Gyimothy, T. (2004). The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. (pp. 41-49). Brno, Czech Republic: Seventh International Conference on Text, Speech and Dialogue (TSD 2004).

Jurafsky, D., & Martin, J. (2009). *Speech and Language Processing.* New Jersey: Prentice Hall, Upper Saddle River, ISBN-10: 9780131873216.

Mani, I., & Shiffman, B. (2005). Temporally Anchoring and Ordering Events in News. In *Time and Event Recognition in Natural Language.* Amsterdam: John Benjamins.

Moldovan, D., Clark, C., & Harabagiu, S. (2005). Temporal Context Representation and Reasoning. *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-2005* (pp. 1099-1104). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sauri, R., Knippen, R., Verhagen, M., & Pustejovsky, J. (2005). Evita: A Robust Event Recognizer for QA Systems. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, (pp. 700-707).

Subecz, Z. (2014). Detection and Classification of Events in Hungarian Natural Language Texts. *Proceedings of the 17th International Conference, TSD 2014* (pp. 68-75). Brno, Czech Republic: Springer Lecture Notes in Computer Science Volume 8655.

Subecz, Z. (2015a). Automatic Labeling of Semantic Roles with a Dependency Parser in Hungarian Economic Texts. *Springer* (pp. 261-272). Brno, Czech Republic: Springer, 18th International Conference on Text, Speech and Dialogue, TSD 2015.

Subecz, Z. (2015b). Szemantikus szerepek automatikus címkézése függőségi elemző alkalmazásával magyar nyelvű gazdasági szövegeken. *XI. MAGYAR SZÁMÍTÓGÉPES NYELVÉSZETI KONFERENCIA* (pp. 95-106). Szeged: Szegedi Tudományegyetem.

Subecz, Z. (2016). Automatic Detection of Nominal Events in Hungarian Texts with Dependency Parsing and WordNet. *Information and Software Technologies, 22nd International Conference, ICIST 2016* (pp. 580-592). Druskininkai, Lithuania: Springer.

Subecz, Z. (2017a). Event Detection in Hungarian Texts with Dependency and Constituency Parsing and WordNet. *Informatics 2017, IEEE 14th International Scientific Conference on Informatics* (pp. 365-371). Poprad Slovakia: IEEE Xplore.

Subecz, Z. (2017b). Főnévi események automatikus detektálása függőségi elemző és WordNet alkalmazásával magyar nyelvű szövegeken. *XIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 13-24). Szeged: Szegedi Tudományegyetem.

Subecz, Z., & Csák, É. (2014). Igei események detektálása és osztályozása magyar nyelvű szövegekben. *X. Magyar Számítógépes Nyelvészeti Konferencia*, (pp. 237–247). Szeged.

Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P. (2003). Using Predicate-Argument Structures for Information Extraction. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 8-15). Sapporo, Japan: Association for Computational Linguistics.

Szarvas, G., Farkas, R., & Kocsor, A. (2006). A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. *The Ninth International Conference on Discovery Science on Discovery Science* (pp. 267-278). Barcelona, Spain: Springer Verlag Berlin, Heidelberg, LNAI 4265, ISBN:3-540-46491-3.

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., & Csirik, J. (2010). Hungarian Dependency Treebank. *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (pp. 1855-1862). Valletta, Malta.: Springer.

Zsibrita, J., Vincze, V., & Farkas, R. (2013). magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. *Proceedings of RANLP-2013, International Conference on Recent Advances in Natural Language Processing* (pp. 763–771). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA.