Detection and Classification of Events in Hungarian Natural Language Texts

Zoltán Subecz

College of Szolnok, Department of Economic and Analysis Methodology,
Tiszaligeti sétány 14, 5000 Szolnok, Hungary
subecz@szolf.hu

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. In this study we introduce new methods for this application area. According to our best knowledge ours is the first result for detection and classification of verbal and infinitival events in Hungarian natural language texts. Evaluating them on test databases, our algorithms achieved competitive results as compared to the current English results.

Keywords: Information extraction, Event detection, Event classification

1 Introduction

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.)* Several events are expressed with two words, e.g. *döntést hoz (make a decision)*, these require distinct treatment. Some studies have dealt with multiword verbal expressions in detail [10,11], we utilized their results here.

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.

In our study we introduce new methods for this 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. Evaluating them on test databases, our algorithms achieved competitive results as compared to the current English results.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 68–75, 2014.

2 Related Work

Several papers are concerned with event detection. The most studies focused only on particular events (e.g. business events). In our present work we dealt with the detection and classification of **all** verbal and infinitival events. Most studies engaged in detection and classification of verbal events for only English texts.

Bethard [1] detected events with statistical features. He took into account multi-word expressions too. The model achieved an F-measure of 88.3 for detection and 70.7 for classification.

Llorens et al. [6] applied a CRF model with the application of semantic rules for event detection and classification. The model achieved an F-measure of 91.33 for detection and 73.51 for classification.

Marsaic [7] focused only on verbal event detection and classification. For detection the model achieved an F-measure of 86.49.

The previous studies were designed for the English language.

Bittar [2] detected events in French texts, and achieved an F-measure of 88.8.

Kata et al. [5] applied a clustering algorithm for Hungarian verbs.

Subecz et al. [9] detected and classified verbal events in Hungarian texts, but their methods and results were simplified, not well-worked-out, and was published only in Hungarian language.

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 [8] for the semantic characterization of the examined words, and we disambiguated the polysemic inspected words with the Lesk algorithm [4].

3 The Corpus and Applied Software Packages

In our application we used one part of the Szeged Corpus [3], 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.

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).

The J48 decision tree algorithm of the Weka 3 data mining suite was employed for machine learning. For the linguistic processing of Hungarian texts the Magyarlanc 2.0 [12] toolkit was used.

4 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.

4.1 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.
- **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. [4]: 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.

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.

5 Results – Event Detection

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

Our first baseline method achieved an F-measure of 79.45, the second one 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.

We examined the efficiency of the particular feature groups with an ablation analysis. In this case the particular feature groups were left out from the whole feature set, and we trained on the basis of the residual features. The results can be found in Table 1. According to the results the Semantic and Frequency 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.

Then we tested our model on verbs and without the rule based method. We got an F-measure of 94.75 with focusing only on verbs. We got an F-measure of 95.20 without the rule based method. Henceforward the rule based method was used together with focusing on verbs and infinitives.

We examined the model's performance on each subcorpus by randomly splitting the particular subcorpus into training/evaluation sets in a 9/1 ratio. 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 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*	−0.03
Syntactic	94.80	95.99	*95.39*	−0.09
Semantic	94.63	96.06	***95.34***	**−0.14**
Frequency	92.70	96.26	***94.45***	**−1.03**

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*

5.1 Additional Experiments for Event Detection

These experiments did not improve the results for event detection.

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 an F-measure of 95.33 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: *OBJ-book*. This extended model achieved an F-measure of 95.39 with 10 fold cross validation.

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.

7 Results – Event Classification

The following experiments on event detection 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 we selected 11 frequent verbs that denote movement and 16 frequent verbs that denote communication events. The model treated only these events as 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

We examined the significance of the particular feature groups with an ablation analysis. In this case the particular feature groups were left out from the whole feature set, and we trained on the basis of the residual features. The results can be found in Table 3. According to the results, the Semantic and Frequency features proved to be the most useful ones. According to the average differences, the best results were achieved without the Morphological features, therefore our further experiments were performed without them.

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

Left out features	Action	Occurrence	Movement	Communication	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
Morphological	87.10	73.66	68.72	82.34	+0.4/+0.23/+0.21/+0.77
Syntactic	85.58	73.54	68.54	80.74	−1.48/+0.11/+0.03/−0.83
Semantic	**86.2**1	**72.52**	**66.02**	**80.22**	**−0.85/−0.91/−2.49/−1.35**
Frequency	**85.58**	**71.16**	**60.76**	**79.93**	**−1.48/−2.27/−7.75/−1.64**

We examined the model's performance on each subcorpus by randomly splitting the particular subcorpus into training/evaluation sets in a 9/1 ratio. 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 - F-measure - Event classification

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

7.1 Additional Experiments for Event Classification

In the next two paragraphs we marked bold the improved results 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-teache*r. This extended model achieved F-measures of **A: 87.63; O: 74.04; M: 68.92;** C: 81.69 with 10 fold cross validation.

8 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. According to our best knowledge ours is the first result for detection and classification of verbal and infinitival events in Hungarian natural language texts. We tested the model's feature set with an ablation analysis, then the model's performance on 5 subcorpora. Evaluating them on test databases, our 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 for the four classifications.

References

1. Bethard, S.J.: Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach (2002) Computer Science. Boulder, CO, University of Colorado
2. Bittar, A.: Annotation of Events and Temporal Expressions in French Texts, ACL-IJCNLP '09 Proceedings of the Third Linguistic Annotation Workshop, 2009, pp. 48–51
3. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Sojka, P. et al. (Eds.) Proc. of the Seventh International Conference on Text, Speech and Dialogue (TSD 2004), Brno, Czech Republic, 8–11 September, pp. 41–49 (2004)
4. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall, Upper Saddle River, NJ, 2000.
5. Kata, G., Héja, E.: Clustering Hungarian verbs on the basis of complementation patterns. Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop, 2007, pp. 91–96
6. Llorens, H., Saquete, E., Navarro-Colorado, B.: TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles, COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 725–733
7. Marsic, G.: Temporal processing of news: annotation of temporal expressions, verbal events and temporal relations. PhD thesis, University of Wolverhampton, 2011
8. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., 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), Szeged, University of Szeged (2008) 311–320
9. Subecz, Z., Nagyné, Cs.É.: Igei események detektálása és osztályozása magyar nyelvű szövegekben. X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2014, p. 237–247
10. Vincze, V.: Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban. VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2009, pp. 390–393
11. Vincze, V., Zsibrita, J., Nagy, T.I.: Dependency Parsing for Identifying Hungarian Light Verb Constructions. In: Proceedings of International Joint Conference on Natural Language Processing 2013, pp. 207–215.
12. Zsibrita,J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2013, pp. 368–374