Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents

Jan Kocoń(✉) and Michał Marcińczuk

Department of Computational Intelligence, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
{jan.kocon,michal.marcinczuk}@pwr.edu.pl

Abstract. In this article we present the result of the recent research in the recognition of events in Polish. Event recognition plays a major role in many natural language processing applications such as question answering or automatic summarization. We adapted TimeML specification (the well known guideline for English) to Polish language. We annotated 540 documents in Polish Corpus of Wrocław University of Technology (KPWr) using our specification. Here we describe the results achieved by Liner2 (a machine learning toolkit) adapted to the recognition of events in Polish texts.

Keywords: Information extraction · Event recognition · Polish wordnet

1 Introduction

Event recognition is one of the information extraction tasks. In the general understanding an *event* is anything what takes place in time and space, and may involve agents (executor and participants). In the context of text processing, event recognition relies on identification of textual mentions, which indicate events and describe them. In the literature there are two main approaches to this task: *generic* and *specific*. The *generic* approach assumes a coarse-grained categorization of events and is focused mainly on recognition of event mentions (textual indicators of events). Such approach is exploited in the TimeML guideline [1]. In turn the *specific* approach is focused on a detailed recognition of some predefined events including all components which describe them. This approach assumes that there is a predefined set of event categories with a complete description of their attributes. For example the ACE English Annotation Guidelines for Events [2] defines a *transport* as an event which "occurs whenever an ARTIFACT (WEAPON or VEHICLE) or a PERSON is moved from one PLACE (GPE, FACILITY, LOCATION) to another". The *specific* approach is a domain- or task-oriented for dedicated applications. In our research we have focused on the *generic* approach as it can be utilized in any domain-specific task. According to our best knowledge this is the first research on automatic recognition of generic events for Polish.

P. Sojka et al. (Eds.): TSD 2016, LNAI 9924, pp. 12–19, 2016.
DOI: 10.1007/978-3-319-45510-5_2

2 Event Categories

In our research we have exploited the coarse-grained categorization of events defined in TimeML Annotation Guidelines Version 1.2.1 [1]. TimeML defines seven categories of events, i.e., *reporting*, *perception*, *aspectual*, *intentional action*, *intentional state*, *state* and *occurrence*. We use a modified version of the TimeML guidelines[1]. One of the most important changes is the extension of the *occurrence* category. According to TimeML *occurrence* refers only to specific temporally located events. Instead, we use an *action* category which include also generics — actions which refer to some general rules (for example, "Water boils in 100°C"). We argue that the distinction between specific and generic actions is much more complex task than the identification of action mentions and may require discourse analysis. Also the event generality applies to the other categories of events as well. Thus, it should be treated as an event's attribute rather than a category. The other important modification, comparing the original TimeML guidelines, is the introduction of the *light predicates* category. This category is used to annotate synsemantic verbs which occur with nominalizations. This type of mentions does not contain enough semantic information to categorize the event. They carry only a grammatical and very general but sufficient lexical meaning which can be useful in further processing. A similar category was introduced by [4] in their research on event recognition for Dutch. The remaining categories have the same definition as in the TimeML guidelines. The final set of event categories contains: *action*, *reporting*, *perception*, *aspectual*, *i_action*, *i_state*, *state* and *light predicate*.

3 Data Sets

3.1 Corpus

In the research we used 540 documents from the Corpus of Wrocław University of Technology [5] which were annotated with events by two linguists according to our guidelines (see Sect. 2). We prepared two divisions for the purpose of the evaluation, which are presented in Table 1.

3.2 Inter-annotator Agreement

The inter-annotator agreement was measured on randomly selected 200 documents from KPWr. We used the positive specific agreement [6] as it was measured for T3Platinum corpus [7]. Two linguists annotated the randomly selected subset. We calculated the value of the positive specific agreement (PSA) for each category. The results are presented in Table 2.

According to [7] the best quality of data was achieved for TempEval-3 platinum corpus (T3Platinum, which contains 6375 tokens) and it was annotated and reviewed by the organizers. Every file was annotated independently by

[1] The comprehensive description of the modified guidelines is presented in [3].

Table 1. Description of two divisions of 540 documents from KPWr annotated with events. The first division is used to establish a baseline (see Sect. 6.1) and the second division is used to evaluate the impact of the generated dictionary features added to the baseline feature set for the result of the events recognition (see Sect. 6.2).

Division	Data set	Documents	Part of whole [%]
1	train1	270	50
	test1	135	25
	tune1	135	25
2	train2_p1	216	40
	train2_p2	216	40
	test2	108	20

Table 2. The value of positive specific agreement (PSA) calculated on the subset of 200 documents from KPWr, annotated independently with events by two domain experts. *A and B* means all annotations in which annotators A and B agreed. *Only A* is the number of annotations made only by annotator A and *only B* – the number of annotations made only by annotator B.

Category	A and B	Only A	Only B	PSA [%]
action	5268	1042	771	85.32
aspectual	100	31	23	78.74
perception	53	58	15	59.22
reporting	64	34	27	67.72
i_action	86	243	18	39.72
i_state	409	123	112	77.68
state	681	281	335	68.86
light predicate	84	119	61	48.28
$\sum$	6745	1931	1362	80.38

atleast two expert annotators and a third was dedicated to adjudicating between annotations. The result of overall T3Platinum inter-annotator positive specific agreement (PSA) at the level of annotating of events only was 0.87 and for the agreed entity set (exact matches) it was 0.92. It means that annotators agreed at the type of annotation at 0.92 for the annotations, which extents were agreed at 0.87, which for the task of manual annotation of both boundaries and event category is approximately 0.80. In our case for 200 randomly selected documents the PSA value achieved for the task of manual annotation of both boundaries and event categories was also 0.80. Unfortunately, for the corpus presented in [7] we see only the overall result for all event categories and we cannot compare the results for each category separately.

4 Generating of Event Dictionaries

The underlying hypothesis of this approach is that generalisation of specific words (*event mentions* in our case) in a subset of documents from a corpus allows to locate synsets in a wordnet, for which we can reconstruct dictionaries, which describe the observed phenomenon and allows to distinguish between different semantic categories of words (in our case — *event categories*) observed in the same set of documents. The algorithm consists of the following steps:

Construction of the helper graph — for each synset w from wordnet synsets W we add a subset of child lemmas C_w, which contains all lexical units from the synset and lexical units of its all hyponyms.

Building the corpus category vector — for the subset S of documents from the corpus and for the number of observed categories T (in our case 8 categories of events + *0 category* for words which do not indicate any event) we build $|T|$ vectors V. For each vector V^t describing the category $t \in T$, the length $|V^t|$ is equal to the number of words from the subset S and the value on n-th position (which represents n-th word in S) equals 1 if word S_n belongs to category t, 0 otherwise.

Building the corpus synset vector — for each $(w, C_w) \in W$ we build a vector A_w. The length $|A_w|$ is equal to the number of words from subset S and the value on n-th position (which represents n-th word in S) equals 1 if word $S_n \in C_w$, 0 otherwise.

Calculating the Pearson's correlation — for each $w \in W$ and each $t \in T$ we calculate the value of a Pearson's correlation $P_w^t = pearson(V^t, A_w)$.

Selection of the best nodes in hyponym branches — for each $t \in T$ we selected only these synsets from W, for which the value of P_w^t was the highest and the lowest in each hyponym branch. B_+^t is the subset of synsets and their child lemmas with the highest positive Pearson's correlation values in each hyponym branch of WordNet, and B_-^t is the subset of synsets and their child lemmas with the lowest negative Pearson's correlation values in each hyponym branch of wordnet. The whole process can be also driven with a given threshold p, which means the minimum absolute value of calculated Pearson's correlation to add a synset to B_+^t or B_-^t. In our experiments we used $p = 0.001$.

Selection of the best B_+, B_- subsets — we built a method for each $t \in T$ to combine the best nodes in hyponym branches to construct a pair of subsets (L^t, H^t) where $L^t \in B_-^t$ and $H^t \in B_+^t$ of the best nodes for which the value of Pearson's correlation calculated between V^t and a modified corpus synset vector M^t built on a pair (L^t, H^t) would be the highest. A length of a modified vector $|M^t|$ is equal to a number of words from subset S and the value on n-th position (which represents n-th word in S) equals 1 if word $S_n \in H^t \vee S_n \notin L^t$ and 0 otherwise. Constructing of (L^t, H^t) is iterative and requires to construct only H^t first. To do that in each step we try to add $b \in B_+^t$ to H^t, recalculate M^t and check if $pearson(V^t, M^t)$ is higher. In each step we find $b \in B_+^t$ which gives the highest gain to the value of

$pearson(V^t, M^t)$ and we add b to H^t and we remove b from B^t_+. We do that as long as we have a positive gain. Then, having H^t, we do the same with B^t_- and L^t.

Generating of dictionaries — for each $t \in T$ we generate separate positive D^t_+ and negative D^t_- dictionaries for H^t and L^t:
$\forall t \in T \forall (w, C_w) \in H^t, D^t_+ = D^t_+ \cup C_w$
$\forall t \in T \forall (w, C_w) \in L^t, D^t_- = D^t_- \cup C_w$

5 Recognition

Many state of the art systems which recognize events use supervised sequence labeling methods, mostly Conditional Random Fields (CRFs) [8]. Recent workshops about the comparison of event recognition systems like TempEval-2 and TempEval-3 [9] show a shift in the state-of-the-art. Currently the recognition of events is done best by supervised sequential classifiers instead of rule-engineered systems [9]. The best machine learning system reported by UzZaman [9] — TIPSem [10] — utilizes CRFs in the recognition of events.

Our approach is based on the *Liner2* toolkit[2] [11], which uses CRF++[3] implementation of CRF. This toolkit was successfully used in other natural language engineering tasks, like recognition of Polish named entities [11,12] and temporal expressions [13].

6 Evaluation

6.1 Feature Selection and Baseline Features

In recognition, the values of features are obtained at the token level. As a *baseline* we used a result of the selection of features from the *default* set of features available in the Liner2 tool. It contains 4 types of features: morphosyntactic, ortographic, semantic and dictionary. We described the *default* set of 46 features in the article about the recognition of Polish temporal expressions [13].

The detailed description of the selection process is available in [13]. Table 3 presents the result of the feature selection from the *default* set of 46 features, based on the F_1-score of 10-fold cross-validation on *tune1* data set.

Table 4 presents the comparison of average F_1-score for all event categories achieved on *train1* and *test1* data sets and for two feature sets: *default* and *baseline*.

We analyzed the statistical significance of differences between two feature sets on two different data sets. To check the statistical significance of F_1-score difference we used paired-differences Student's t-test based on 10-fold cross-validation with a significance level $\alpha = 0.05$ [14]. Differences are not statistically significant for both data sets, but the reduction of a feature space is from 46 to only 6 features, which compose a *baseline* set of features for the further evaluation.

[2] http://nlp.pwr.wroc.pl/en/tools-and-resources/liner2.
[3] http://crfpp.sourceforge.net/.

Table 3. Result of the feature selection for Polish events recognition used in this work as a *baseline*. Used measure: average *exact match* F_1-score of 10-fold cross-validation on *tune1* set. Initial set of features: default 46 Liner2 features.

Iteration	Selected feature	F_1 [%]	Gain [pps]
1	class	62.13	62.13
2	hypernym-1	73.23	11.10
3	top4hyper-4	74.68	1.45
4	prefix-4	75.26	0.58
5	struct	75.85	0.59
6	synonym	76.30	0.45

Table 4. Comparison of results (F_1-score) achieved on two data sets (*train1* – 10-fold cross-validation on *train1* set; test1 – model is trained on *train1* set and evaluated on *test1* set) and two feature sets (*default* – 46 default features available in Liner2; *baseline* – result of the feature selection on *default* feature set and *tune1* data set).

Set	Default [%]	Baseline [%]
train1	77.47	77.53
test1	78.90	78.34

6.2 Baseline with Dictionary Features

We generated two sets of dictionaries for each part of *train2* set (these parts are fully separated). Dictionaries were created using the plWordNet [15] — the largest wordnet for Polish. We used dictionary features generated on *train2_p1* to evaluate the model on *train2_p2* data set, and then we used dictionary features generated on *train2_p2* to evaluate the model on *train2_p1* data set. The last two models (first trained on *train2_p1* and second trained on *train2_p2*) were evaluated using *test2* data set.

Table 5. Comparison of results (F_1-score) achieved on two **Part**s of *train2* data set: *p1* and *p2*. These data sets were also dictionary **Source**s for **B+dict** feature set, to compare results with **B**aseline feature set. We performed two types of **Eval**uation: *CV* (10-fold cross-validation on a part of *train2* set) and *test2* (the model is trained on a part of *train2* set and the evaluation is performed on a *test2* set).

Part	Eval.	Source	B [%]	B+dict [%]
p1	CV	p2	77.20	79.67
p2	CV	p1	76.92	78.92
p1	test2	p2	77.65	79.87
p2	test2	p1	77.81	79.82

We performed 4 tests to evaluate the impact of generated dictionary features added to the baseline feature set for the result of the events recognition. The result is presented in Table 5. We see that in each test we achieved better results with the set of features extended with dictionaries. We analyzed the statistical significance of differences between these results for each test. To check the statistical significance of F_1-score difference we used paired-differences Student's t-test based on 10-fold cross-validation with a significance level $\alpha = 0.05$ [14]. All differences are statistically significant.

7 Conclusions

In Table 6 we present the comparison of detailed results for each event category, achieved on both parts of *train2* data set (as a sum of True Positives, False Positives and False Negatives of 10-fold cross-validation on *train2_p1* and 10-fold cross-validation on *train2_p2*).

Table 6. Comparison of detailed results for each event **Category** achieved on both parts of *train2* data set (the result is the sum of TP, FP and FN of 10-fold cross-validation on *train2_p1* and *train2_p2*). **Baseline+dict** variant is a set of **Baseline** features extended with dictionary features. The last column shows the value of **PSA** (positive specific agreement), described in Sect. 3.2.

Category	Baseline			Baseline+dict			PSA
	P [%]	R [%]	F [%]	P [%]	R [%]	F [%]	[%]
action	80.05	83.14	81.57	82.49	83.87	83.18	85.32
aspectual	93.00	39.08	55.03	87.58	59.24	70.68	78.74
i_action	63.97	27.87	38.82	63.56	40.92	49.79	59.22
i_state	85.82	75.08	80.09	85.19	77.56	81.20	67.72
light_predicate	50.00	11.03	18.07	56.76	15.44	24.28	39.72
perception	96.55	23.14	37.33	85.90	55.37	67.34	77.68
reporting	73.91	44.24	55.35	71.13	51.30	59.61	68.86
state	70.81	50.19	58.74	68.10	62.17	65.00	48.28
$\sum$	79.54	74.75	77.07	80.88	77.82	79.32	80.38

We see that adding dictionary features statistically significantly increased the result of events recognition. Detailed analysis performed on separate event categories showed that the major improvement can be observed with categories which are underrepresented in corpus and for which the PSA value was smaller. For models with dictionary features the F-measure is more close to PSA values for all categories of events. Dictionary features increased the values of both precision and recall.

Acknowledgments. Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References

1. Saurí, R., Littman, J., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML Annotation Guidelines, Version 1.2.1 (2006)
2. LCD: ACE (Automatic Content Extraction) English Annotation Guidelines for Events (Version 5.4.3). Technical report, Linguistic Data Consortium (2005)
3. Marcińczuk, M., Oleksy, M., Bernaś, T., Kocoń, J., Wolski, M.: Towards an event annotated corpus of Polish. Cogn. Stud. Études Cogn. **15**, 253–267 (2015)
4. Schoen, A., van Son, C., van Erp, M., van der Vliet, H.: NewsReader document-level annotation guidelines - Dutch. NWR-2014-08. Technical report, VU University Amsterdam (2014)
5. Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., Wardyński, A.: WUTC: towards a free corpus of Polish. In: Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, 23–25 May 2012 (2010)
6. Hripcsak, G., Rothschild, A.S.: Agreement, the F-measure, and reliability in information retrieval. J. Am. Med. Inform. Assoc. **12**, 296–298 (2005)
7. UzZaman, N., Llorens, H., Allen, J.F., Derczynski, L., Verhagen, M., Pustejovsky, J.: TempEval-3: evaluating events, time expressions, and temporal relations. CoRR abs/1206.5333 (2012)
8. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco (2001)
9. UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., Pustejovsky, J.: SemEval-2013 task 1: TEMPEVAL-3: evaluating time expressions, events, and temporal relations, Atlanta, Georgia, USA, p. 1 (2013)
10. Llorens, H., Saquete, E., Navarro, B.: TipSEM (English and Spanish): evaluating CRFs and semantic roles in TempEval-2. In: Association for Computational Linguistics, pp. 284–291 (2010)
11. Marcińczuk, M., Kocoń, J., Janicki, M.: Liner2 – a customizable framework for proper names recognition for Polish. In: Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., Niezgódka, M. (eds.) Intelligent Tools for Building a Scientific Information. SCI, vol. 467, pp. 231–254. Springer, Heidelberg (2013)
12. Marcińczuk, M., Kocoń, J.: Recognition of named entities boundaries in Polish texts. In: ACL Workshop Proceedings (BSNLP 2013) (2013)
13. Kocoń, J., Marcińczuk, M.: Recognition of Polish temporal expressions. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2015) (2015)
14. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. **10**, 1895–1923 (1998)
15. Maziarz, M., Piasecki, M., Szpakowicz, S.: Approaching plWordNet 2.0. In: Proceedings of the 6th Global Wordnet Conference, Matsue, Japan (2012)

http://www.springer.com/978-3-319-45509-9

Text, Speech, and Dialogue
19th International Conference, TSD 2016, Brno , Czech Republic, September 12-16, 2016, Proceedings
Sojka, P.; Horák, A.; Kopeček, I.; Pala, K. (Eds.)
2016, XXI, 550 p. 109 illus., Softcover
ISBN: 978-3-319-45509-9