Events Extractor for Polish in a Semantics-Driven Mode

Jolanta Cybulka and Jakub Dutkiewicz

Poznań University of Technology
3B Piotrowo Street, Poznań, Poland
{jolanta.cybulka,jakub.dutkiewicz}@put.poznan.pl

Abstract

In this paper the events extractor for Polish is presented. The considered tool has two components: the first one is domain-dependent and serves to generate extraction templates. The process of generation is driven by semantics of a domain represented by a well-founded ontology. The second part is linguistic and domain-independent and may be used whenever templates are supplied, not necessarily via the generator. We checked the quality performance of our generator on a basis of a case study.

Keywords: event extraction, valence frames, semantics-driven event extraction for Polish

1. Introduction

Event extraction (EE) is a kind of information extraction (IE) and relies on obtaining from free-texts a predefined types of facts concerning events: we aim to detect what happened and what were the arguments/parameters of it. In such case the extraction is domain-dependent for we need the explicit representation knowledge concerning events in some domain of interest. Once the domain semantics is established one may create event extraction templates. In our work we do it automatically generating a knowledge frame and templates on the basis of a well-founded ontology (see section 2). Thus we use our extractor in a semantics-driven mode as opposite to the simple syntactic mode where the templates are manually created. The generated templates are parsed by a linguistic domain-independent part of the extractor. We briefly describe the extraction process and introduce the basic data structures used in the process (see section 3). We evaluate the extraction with a case study with a specific regard to the data structures (see section 4).

2. Generation of extraction templates

The semantics-driven extraction of events from free-texts requires having a domain knowledge to be explicitly specified. To us this need may be ideally fulfilled by using a well-founded ontology. We use a capsular c.DnSPL ontology (Cybulka, 2015) which is based on the foundational ontological pattern of *constructive descriptions and situations* of (Gangemi et al, 2007). This pattern provides a general view on reality, every domain is specified as a capsule created according to that pattern. An important feature, from the EE point of view, of an ontology is the existence of *perdurants* (i.e. entities "that happen") and connected with them thematic roles as concepts. This forms a framework to semantically specify the events. Considering the ontological nature of a kidnapping event we see that not only the *perdurant* should be specified but also its arguments such as: a perpetrator, victim, location in time and place, and optionally a beneficiary, source and target places, manner, result etc. These parameters are modelled by using *thematic roles*, respectively a role of an: *agent*, *patient-object*, *location*, *patient-beneficiary*, *ablative location*, *allative location*, *manner*, *result*. The semantics specified in the ontology allows us to filter the inadequate facts, but it may be "dressed" in many syntactic forms. The question arises, how to provide them? To specify the possible syntactic expressions of the underlined semantics we use role approximations proposed in (Jaworski and Przepiórkowski, 2014). With the help of them we built equivalents between thematic roles (concepts in the ontology) and syntactic *theta-roles* (parameters of verbs in valence structures). In (Cybulka et al., 2015) we present a method of a knowledge frame and template generation and its supporting tool.

3. Specification of event extraction process

As it was mentioned, event extraction is a process of textual data analysis with the goal of knowledge acquisition, in the context of the domain semantics specified by an ontology. The extractor uses specific data structures and tools. The architecture of the extractor is shown in Figure 1.

There are three main data structures used in our EE process:

- *Event Frames*
- *Event Templates*
- *Event Instances*.

Event Frames and *Event Templates* are domain-dependent and may be generated from the ontology or may be manually created. Thus, they are to be instantiated before the extraction process commences, while *Event Instances* are generated by the extractor. *Event Frame* is a semantic description of the event. A singular frame specifies a concept that is hidden under the anchoring word in the event, such as *kidnapping*, as well as pairs: a thematic role and its semantics (semantic classes), such as: *an agent role – a human concept*. Various lexicalizations of the concepts should be necessarily included in the ontology. *Event Templates* specify the valence structure of sentences, which are supposed to be the crucial input source of the extraction. For example, we may specify that the role of an *agent* should be a nominal phrase while the *patient* – a noun phrase with the main noun in the accusative case. *Event instances* are results of EE and describe the details of the instantiated extracted events.

The extraction starts with the tokenizing and tagging process. Tokenizing is handled by the TaKIPI tool

(Piasecki, 2007). Then, the tokens are chunked into phrases.

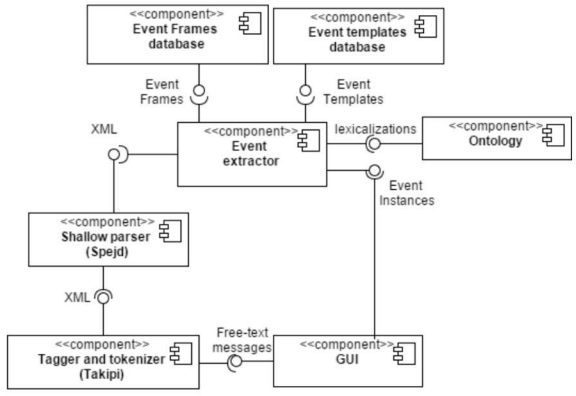

Fig. 1: UML's diagrammatic view of the event extractor

The following parts of the EE process is executed with the Spejd shallow parsing tool for Polish (Przepiórkowski, 2008) according to the specified grammar. The chunking process generates three main types of chunks: noun phrases (NP), verb phrases (VP) and prepositional phrases (PP), as well as several technical chunks, such as: complex sentence separator, (which implies that the potential nominal phrase of the latter homogenous part of the sentence is equal to the nominal phrase of the first homogenous part of the sentence (CSnf)) and complex sentence separator (which implies that the potential nominal phrase of the latter homogenous part of the sentence is equal to the word, which appears directly before the separator (CSf)). The chunks consist of tokens, as well as the identifier of the semantically main word in the chunk and the syntactically main word in the chunk (Figure 2), the exemplary sentence is taken from NJKP[1].

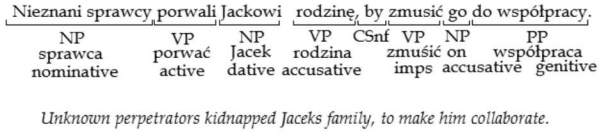

Figure 2: Chunking of the sentence

Once the chunking process is done, the generated XML output is serialized into the C# application. Data is incorporated into the extractor mechanism. The extractor compares the structure of the serialized data with all of the *Event Frames* and *Event Templates*. If the sentence happens to contain a word that is lexicalized by one of the anchoring words within all of the *Extraction Frames*, the sentence is passed for the further examination. In the next step, the extractor tries to match the valence structure with all of the corresponding Event Templates. On top of that, all of the thematic roles must correspond to the lexicalizations. Details of this process are discussed in the (Dutkiewicz et al., 2014). The final result of the extraction is visible in Figure 3. The ontology is the core piece of the extraction process. It contains all the data related to the events semantics, as well as the entirety of the lexicalizations. If the sentence happens to have a word, which does not appear within the lexicalizations, the word is passed to the supporting NER system. If the word is still not recognized, it is treated as if it was valid and corresponding to the correct class. The ontology currently contains 31 event classes and over 200 various classes, which may correspond to the arguments of the event. Each class is supported by a set of lexicalizations expressed in both Polish and English languages. Based on the ontology, it is possible to generate all of the *Event Frames* and Event Templates. The generation process is introduced in (Cybulka et al., 2015). In the next section, we are going to introduce an alternative, manual process of generation of *Event Frames* and *Event Templates*.

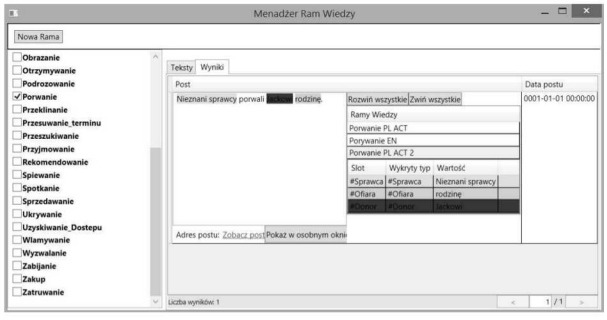

Fig.3: Extraction process results

4. Event extraction templates at work - a case study

The process of automatic *Event Frames* and *Event Templates* generation is described in (Cybulka et al., 2015). In this paper we will go through the procedure of manual template creation on the basis of the *kidnapping* event (the valence structures of this verb are specified for Polish in Walenty[2]). The most basic *Event Frame* for the *kidnapping* event should consist of three thematic roles:

- Event Frame name: *kidnapping event*
- Anchoring word: *kidnapping*
- Agent class: *human* (a concept that acts for a *Person* in the c.DnSPL ontology)
- Patient class: *human*
- Beneficiary class: *human*

Figure 4 depicts the screen of a tool supporting the manual creation of an *Event Frame*. The tool allows one to set up all of the mentioned frame elements (*Typ* – a name, *Kotwica* – an anchoring word, *Nazwa slotu* – a thematic role, *dozwoloneTypySemantyczne* – allowed ontological concepts), as well as to choose whether the default thematic roles of time (*Domyślny czas*) and space (*Domyślne miejsce*) should be included into it or not. The exactly one point in time and place is specified by the default roles. It should be noted, that events do not happen in exactly one point in time and that they often are specified inaccurately or relatively, as for example the

[1] NJKP – national corpus for Polish, http://nkjp.pl/index.php?page=0&lang=1.

[2] Walenty, http://zil.ipipan.waw.pl/Walenty.

word "*yesterday*" in the slightly modified NJKP sentence, translated into English:
"Unknown perpetrators kidnapped Jacek's family yesterday."
Using the tool one can also specify a set of lexicalizations, which will be added into the ontological resources.

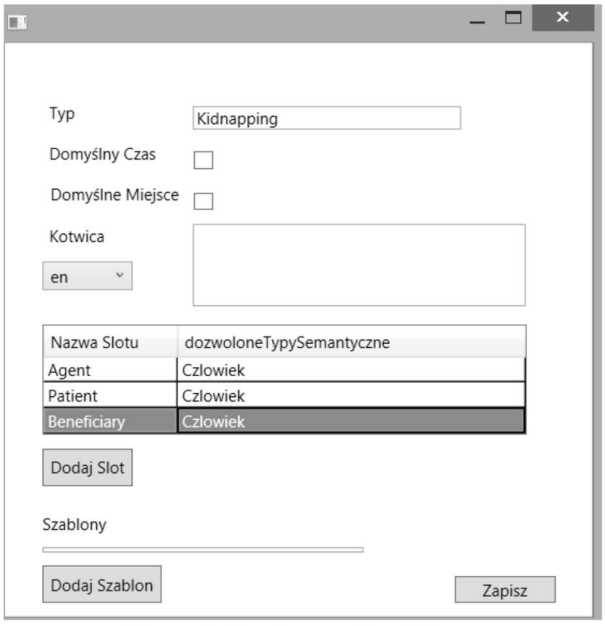

Figure 4. Creating an *Event Frame*

Each *Event Frame* may relate to many *Event Templates*. Let us specify two *Event Templates* for an exemplary *Frame* for kidnapping. At first let us consider the valence structure of the active voice of the anchoring verb:

- Related Frame: *kidnapping event*
- Anchoring word: *active verb phrase*
- Agent: *noun phrase;* Case: *nominative*
- Patient: *noun phrase;* Case: *accusative*
- Beneficiary: *noun phrase;* Case: *dative* – the valence structure in Polish differs from the English one

and then the structure for the passive voice one:

- Related Frame: *kidnapping event*
- Anchoring word: *passive verb phrase*
- Agent: *noun phrase;* Case: *accusative*
- Patient: *noun phrase;* Case: *nominative*
- Beneficiary*: noun phrase;* Case*: dative*

On the basis of this specification the *Event Template* may be created what is shown in Figure 5. The tool allows one to specify all of the previously mentioned features as well as the set of prepositions (*Przyimki*) suitable for the valence structure of verb. Once the operation of creating *Event Templates* and *Event Frames* is done, the data is stored in the ontology.
The defined set of structures enables to extract knowledge from the following sentences:

"Nieznani sprawcy porwali Jackowi rodzinę."
("*Unknown perpetrators kidnapped Jacek's family*.")

"Rodzina Jacka została porwana przez nieznanych sprawców."
("*Jacek's family was kidnapped by unknown perpetrators*.")

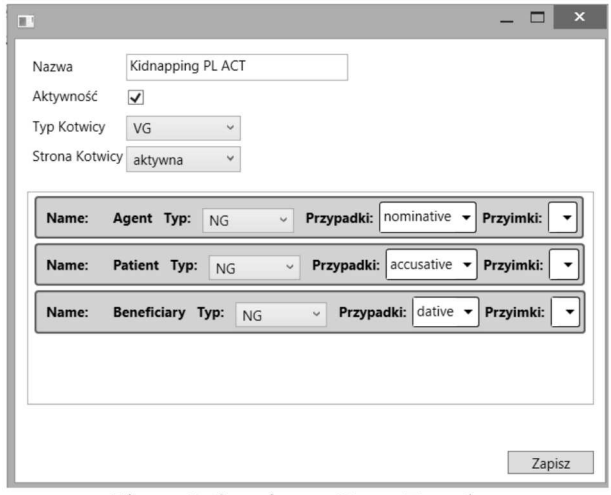

Figure 5. Creating an Event Template

This basic scenario works properly in every similar case of a simple sentence. Let us consider the more complicated (NJKP-based) sentence:

"Jeden z nich w rozmowie lekko odwrócił się od stołu, a wtedy orzechówka porwała z talerzyka ogromny kawał kiełbasy i uciekła."
("*One of them slightly turned back out of table, and then a nutcracker grabbed a huge piece of sausage from the plate and ran away."*)

First of all, the phrase "Jeden z nich", "*One of them*" is never going to be picked up as the *Agent*, because the latter part of the sentence contains a noun phrase with the main noun in nominative case, i.e. "orzechówka" (a *nutcracker*). However, there are several possible outcomes of the extraction process. The results depend both on the vocabulary stored within lexicalizations and on the structure of the event templates. If the words used in the sentence are stored within the set of lexicalizations, the event extractor is not going to make a mistake. Else, if there are no such words stored within lexicalizations, the extractor assumes, that the used word is a correct one. This approach is weak to the homonymic words, such as "orzechówka". The existence of such a word under only one of the semantic classes is most likely going to be confusing in the extraction process. The possible results of extraction from the considered sentence, with regard to the homonymic nature of the word "orzechówka" are listed in Table1.

Table 1. Possible outcomes of the extraction process.

Event Template details	Vocabulary details	Result of the extraction
Patient: either a *human* or a	no "orzechówka" (a *nutcracker*) lexicalization under	no *Event Instance* extracted

food *Agent*: either a *human* or an *animal*	the *animal* class, but there is an "orzechówka" (*whale liqueur*) under another class	
Patient: either a *human* or a *food* *Agent*: either a *human* or an *animal*	"orzechówka" (a *nutcracker*) lexicalization under the *animal* class	*Event Instance* extracted correctly
Agent: a *human*	"orzechówka" (a *nutcracker*) lexicalization under the *animal* (assuming it subsumes a *human*) class	no *Event Instance* extracted
Patient: a *human* *Agent*: an *animal*	"orzechówka" (*whale liqueur*) lexicalization under the *liqueurs* class	*Event Instance* extracted incorrectly*
Agent: a *human*	no "orzechówka" (a nutcracker) lexicalization in the ontology	*Event Instance* extracted incorrectly*

The outcomes denoted with the asterisk (Table 1) incorrectly generate *Event Instances* and should be separated from the others. They illustrate the crucial role of lexicalizations assigned to ontological concepts. This examples also provide an additional information upon further inspection. The acquired in this way information may be used for the vocabulary enrichment. Methods of the vocabulary enrichment are crucial in the process of the accuracy improvement. There is another problem in this particular sentence. The phrase, which describes the patient, consists of nouns in accusative and genitive cases. The phrases in which the other word is in accusative, make it very hard to recognize the semantically most important word. For example:

"Rodzina (nominative – semantically the most important word) Jacka (genitive)"
("*Jacek's Family*")

and

"Kawał (nominative) kiełbasy (genitive – semantically the most important word)"
("*a huge piece of sausage*")

The extractor assumes, that the first word in the phrase is semantically the most important word. In this particular case it does not interfere with the results, however it is a problem, which should be resolved to improve the accuracy of the extraction.

Another example exposes one of the weaknesses of the described approach. The considered sentence is as follows (taken from NJKP):

"Gwiazda wieczoru, Helena Vondračkova wystąpiła dopiero około 00.30 i od razu porwała widownię do tańca."

"The star of the evening, Helena Vondračkova performed only at 00:30 and at once 'kidnapped' the audience to dance."

The idiomatic structures are a problem in our approach. The phrase "*to snatch to dance*" (literally in Polish "*to kidnap to dance*") is a commonly used idiom. In this case, the extraction process generates the *Event Instance* with "*Helena Vondračkova*" as the agent and "*the audience*" as the patient. Both lexicalizations might correctly appear under the human class in the ontology. At first glance, this extraction is correct, however it does not satisfy the intention of the *Event Frame*. The extractor is supposed to find the information about the kidnappings. Generating separate *Event Frames* and ontological classes dedicated for the idiomatic meanings is a solution to this problem. Then, if the extractor did generate an idiomatic *Event Instance*, it should not generate any other *Event Instances*. But such a solution, however, generates a lot of redundancy in the ontology. That problem is to be investigated further.

We have run several experiments with sets of *Event Templates* generated by the approach described in (Cybulka et al., 2015). This approach provided us with the additional information. Creating *Event Templates* manually, usually covers several, most commonly used valence structures, while the before-mentioned approach generates all of the possible templates. There are 53 *Event Templates* for active voice of the kidnapping verb. Let us consider the following sentence (NJKP-based):

"Nieznani sprawcy porwali Krystynę Starczewską z ulcy."
("*Unknown perpetrators kidnapped Krystyna Starczewska from the street.*")

With the set of templates presented at the beginning of the section, the extractor generates an *Event Instance* with two thematic roles. The considered *Event Instance* is partially correct, as it does express the meaning of the sentence, but it omits the ablative role "*from the street*". Obviously, if we were more careful, we would put that role in the template, as well as many other thematic roles (see section 2). However, by doing so, we do not, neither provide nor receive any additional information about the valence structure of the sentence. Creating templates for all of the valence structures is extremely time consuming. With 53 generated templates we created 53 *Event Instances*, out of which 32 cover the ablative role "from the street", and only one *Event Instance* contains 3 thematic roles, which correspond to 3 slots in the *Event Template*. We must note, however, that the time of the extraction process is particularly prolonged by tokenizing and parsing processes, not due to applying many templates.

We have performed similar tests for various sentences with the active and passive voice structures. We achieved similar results in every case. Another example of the *Event Instance* extracted from the complex sentence is depicted in Figure 6. One of the many extracted *Event Instances* covers all of the thematic roles. In this particular case, there is no word "samochód" (a *car*) within the lexicalizations. It makes the extractor assume, that a *car* is a correctly used word, which is unknown to the system. If there was a word "samochód" within the set of lexicalizations, the extraction would be aborted.

"W czwartek miejska policja złapała bandytów, którzy porwali samochód komisarzowi policji w Poznaniu z ulicy"

"On Thursday the city police captured bandits, who sized a car from the street, which belonged to the Poznań police-officer."

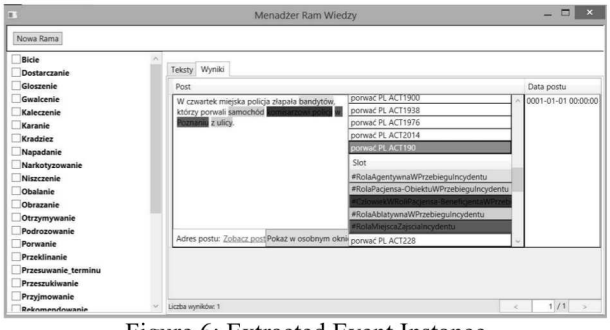

Figure 6: Extracted Event Instance

5. Final remarks

In this paper we provide an extended view on EE process with regard to *Event Frames* and *Event Templates* generation. We firmly believe, that the automated process of template generation provides data which will be useful in the process of (assigned to ontological concepts) vocabulary enrichment, valence structures studies and, ultimately, accuracy improvement in the process of event extraction. We believe, that the research is going to be useful, not only in the event extraction, but also in the synthetic text generation. We are working on the method and tool enabling to assign the proper and rich lexicalizations to ontological concepts. There are at least two problems to be solved: finding a suitable model of lexicalizations and populating the model with relevant lexical data. The considered model may be based on *lemon* linguistic ontology (Lemon – The Lexicon Model for Ontologies). The lexical "seed" for an ontological concept, in the simplest solution, may be given by the user, by means of a dedicated tool. Then, the tool should find proposals of other lexicalizations using a WordNet technology, for example *Słowosieć* (Słowosieć – a large network of words) resource for Polish.

References

Cybulka J. (2015). *The OWL version of c.DnSPL ontology*. Retrieved from: http://users.man.poznan.pl/jolac/PPBW-22-07-2015-inferred-new.owl. Access date: 01-11-2015 (20 MB).

Cybulka, J., Dutkiewicz, J., Żętkowski, M. (2015). Ontology-based Generation of Event Extraction Templates and Frames. Submitted to LTC'2015, 7th Language and Technology Conference, November 27-29 2015, Poznań, Poland.

Dutkiewicz J., Falkowski M., Nowak M., Jędrzejek C. (2014). Semantic Extraction with Use of Frames, Lecture Notes in Computer Science, Springer, ISBN 987-3-319-10888, pp. 208-215.

Gangemi, A., Lehmann, J., Catenacci, C. (2007). Norms and plans as unification criteria for social collectives. In: *Proc. of Dagstuhl Seminar 07122, Normative Multi-agent Systems*, vol. II, pp. 48–87, ISSN 1862-4405.

Jaworski, W., Przepiórkowski, A. (2014). Syntactic Approximation of Semantic Roles. PolTAL 2014. In: A. Przepiórkowski, M. Ogrodniczuk (Eds.): *Advances in Natural Language Processing - 9th International Conference on NLP*, Lecture Notes in Computer Science 8686, Springer, ISBN 978-3-319-10887-2, pp. 193-201.

Lemon – The Lexicon Model for Ontologies. Retrieved from: http://lemon-model.net/index.php. Access date: 01-11-2015.

Piasecki, M. (2007). Polish Tagger TaKIPI: Rule Based Construction and Optimisation. TASK Quarterly 11 No 1–2, 151–167.

Przepiórkowski, A. (2008). Powierzchniowe przetwarzanie języka polskiego, Akademicka Oficyna Wydawnicza EXIT, Warszawa (in Polish).

Słowosieć – a large network of words. Retrieved from: http://plwordnet.pwr.wroc.pl/wordnet/. Access date: 01-11-2015.