Open Relation Extraction for Polish: Preliminary Experiments

Jakub Piskorski
Linguistic Engineering Group
Polish Academy of Sciences
Jakub.Piskorski@ipipan.waw.pl

Abstract

This paper presents preliminary experiments on Open Relation Extraction for Polish. In particular, a variant of a prior-art algorithm for open relation extraction for English has been adapted and tested on a set of articles from Polish on-line news. The paper provides initial evaluation results, which constitute the point of departure for in-depth research in this area.

1 Introduction

While traditional Information Extraction (IE) systems are tailored to the extraction of predefined set of target relations (Appelt, 1999), an Open Information Extraction (OIE) system focuses on the extraction of non predefined, domain-independent relations from texts. The main drive behind the emergence of the OIE paradigm comes from the need to scale IE methods to the size and diversity of the Web (Etzioni et al., 2008).

Analogously to traditional IE, OIE systems deploy either machine-learned extraction patterns, hand-crafted heuristics or a combination of both of them. TEXTRUNNER (Etzioni et al., 2008) was the first OIE system based on a ML approach, where the OIE paradigm was introduced. WOE (Wu and Weld, 2010) is an extension of TEXTRUNNER, where the *Wikipedia* corpus was exploited as training data to boost the coverage. (Etzioni et al., 2011; Fader et al., 2011) introduced REVERB, the first linguistically-lightweight OIE system based on heuristics, which initially identifies verb phrases and light verb constructions that express relations, and subsequently extracts the relations' arguments in the left/right context thereof. (Mausam et al., 2012) and (Del Corro and Gemulla, 2013) are examples of hybrid systems that deploy dependency parsing. Relatively little work has been reported on OIE for non-English languages, e.g., (Gamallo et al., 2012) presents an approach based on dependency parsing and provides evaluation figures for non-English languages. An extensive overview of research and open problems in OIE is provided in (Xavier et al., 2015).

This paper reports on preliminary experiments on developing a scalable linguistically lightweight OIE approach to extraction of arbitrary binary relations from Polish texts. We are particularly interested in extraction of relations from online news. Although the recently reported OIE techniques for extracting binary relations from English texts are advancing rapidly, they might not be directly applicable to languages such as Polish with various phenomena that complicate both IE and OIE tasks (Przepiórkowski, 2007), e.g., relatively free word order, rich morphology (including complex proper noun declension paradigm), syncretism of forms (i.e., single form may fulfill different grammatical functions: subject/object), zero anaphora and existence of pro-drop pronouns. To our best knowledge, the only work on OIE for Polish has been reported in (Wróblewska and Sydow, 2012), where a dependency parsing-based approach to binary relation extraction has been introduced. The main difference of the aforementioned work vs. ours is that the former focuses solely on the extraction of relations that hold between named entities of certain type, whereas in the presented work we do not introduce such limitations. Secondly, we deliberately intend to approach the OIE problem in an incremental manner, i.e., start explorations with as linguistically-poor methods as possible and identify the phenomena/issues that complicate the task at hand most before elaborating more sophisticated solutions, whereas the work described in (Wróblewska and Sydow, 2012) deploys relatively linguistically sophisticated chain of NLP modules, including a dependency parser, which might prohibit applying it

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 34–39,
Hissar, Bulgaria, 10–11 September 2015.

on Web-scale corpora. Finally, although the evaluation results reported in (Wróblewska and Sydow, 2012) are promising, they only refer to a limited number of preselected relation types (e.g., '*born in*), thus making a direct comparison difficult.

2 Simple Relation Extraction for Polish

In order to start explorations on OIE for Polish we developed SREP (Simple Relation Extractor for Polish) that extracts binary relations from free texts in Polish in a form of triples `(arg1,relation,arg2)`, e.g., (*prezydent Komorowski*,*spotkał się z*,*lekarzem*) (president Komorowski, met with, the doctor). SREP is to a large extent a direct adaptation of REVERB (i.e., it borrows the main idea), an open relation extractor for English (Fader et al., 2011). It first identifies candidate relation phrases that satisfy certain syntactic and lexical constraints, and subsequently finds for each such phrase potential NP arguments. In a third (optional) step, a set of generic lexico-syntactic patterns for extracting binary relations is applied to capture specific phenomena and harder-to-tackle constructions in Polish. In case the application of such a pattern covers a larger text span than a text span corresponding to a relation extracted at an earlier stage then the latter is discarded. All identified relation extractions are assigned a confidence score and the ones, for which confidence is higher than a prespecified threshold are returned by the system.

A more detailed description of SREP is given below. In order to create some of the resources described below a *Training Corpus* consisting of ca. 1200 sentences randomly selected from a larger collection of on-line Polish news (*News Corpus*), consisting of 20 MB of text was used.[1].

1. **Pre-processing:** SREP takes as input a sequence of sentences and performs tokenization and morphological analysis thereof. For obtaining part-of-speech information we use *Polimorf* (Woliński et al., 2012), a freely available morphological dictionary for Polish, consisting of circa 6.7 million word forms, including proper names.

2. **Relation Phrase Extraction:** Relation phrase candidates are extracted using a small-scale POS-based regular grammar consisting of 6 patterns, which appeared frequently in the training corpus, e.g., patterns like:

```
1. "nie" V (V)?

2. V (V)? N? "się"? PREP
```

   The second pattern covers for instance the phrase *urodził się w* (was born in) or *zawarł umowę z* (made a deal with). In order to eliminate implausible relation phrases a 'stop' list of phrases[2] is used (e.g., it contains the phrase *niż do* - meaning "bow down to" (something) or "than to", where the second interpretation is more prevalent and is not used to express relations. If any pair of matches overlap or are adjacent then they are merged into a single relation phrase. Each extracted relation phrase is associated with a confidence score that depends on the rule that has triggered the extraction and also other parameters, e.g., length of the extraction. For instance, the second pattern above is less reliable than the first one, hence it is associated with a lower confidence. Confidence score for a given pattern has been computed based on the fraction of 'correct' extractions it produced in the training corpus.

3. **Noun Phrase Recognition:** Analogously to Step 2 NPs are extracted subsequently using 8 POS-based patterns, where each pattern is associated with a confidence score, computed in a similar manner as above, e.g., the pattern `(Adj)+ N (Adj)+`[3] is less reliable than `N+`.

4. **Argument Extraction:** For each relation phrase *rel* identified in Step 2, the nearest noun phrase *X* to the left of *rel* is identified, which is neither a pronoun nor WHO-adverb. Analogously, the nearest noun phrase *Y* to the right of *rel* is identified. In case such *X* and *Y* could be found the system returns (*X*,*rel*,*Y*) triple as an extraction. Each such extraction is assigned a confidence score, which is the product of the confidence of extracting the constituents of the relation triple.

[1] The news articles were gathered using Europe Media Monitor (*emm.newsbrief.eu*), a multilingual news aggregation engine.

[2] This list consist of circa 400 entries and was created based on frequency analysis of application of the aforementioned grammar on the news corpus.

[3] Adjectives may appear in Polish both on the left and right of a noun in a noun phrase.

5. **Application of Lexico-Syntactic patterns (Optional):** A set of 12 generic lexico-syntactic patterns for extracting binary relations is applied optionally at this stage, many of which are intended to cover either more complex constructions or phenomena typical for Polish.[4] In particular, the patterns rely on previously computed relation phrases in Step 2. Some sample patterns are given below (in a simplified form), where `REL` refers to the relation phrases extracted in Step 2.

```
1. NP-1 REL-1 NP-2, "który" REL-2
   NP-2
    -> (NP-1,REL-1,NP-2)
       (NP-2,REL-2,NP-3)

2. NP-1 "to" NP-2 PREP NP-3
   -> (NP-1,NP-2 PREP,NP-3)

3. PREP NP(GEN)-1 REL NP-2
   -> (NP-2,REL-1 PREP, NP(GEN)-1)

4. NP-1 REL NP-2 ("," or CONJ) REL-2
   NP-3
   -> (NP-1,REL,NP-2)
      (NP-1,REL-2,NP-3)
```

The first pattern extracts two relations from a text fragment that includes a relative clause (starting with the word *który* - which), whereas the second pattern covers relations that are not expressed using verbs[5], e.g., *Oborniki to miasto w Wielkopolsce* (Oborniki is a city in Wielkopolska) is covered by this rule and results in the extraction of (*Oborniki*,*miasto w*,*Wielkopolsce*). The third pattern covers a specific construction, in which the relation phrase is not a continuous sequence of tokens, that turns to occur frequently in Polish, e.g., *Do Polski przyjechał prezydent USA* (To Poland has arrived president of USA). Finally, the fourth pattern extracts relations from a particular elliptical construction, e.g., from the sentence *Lech wygrał z Legią i przegrał z Ruchem* (Lech won with Legia and lost to Ruch). Analogously to Step 4, each pattern is assigned a confidence score, which reflects the fraction of correct extractions this pattern triggered on the sentences in the training corpus.

The creation and testing of all underlying linguistic resources mentioned above, i.e., the patterns, took 3-4 days for a single person.

3 Evaluation

Four instances of the algorithm sketched in 2 have been evaluated: SREP (the algorithm without Step 5), SREP-PAT (the algorithm with Step 5), SREP-OV (the algorithm without Step 5, where the text fragments from which relation triples are extracted may overlap, e.g., two relations are extracted from the same text fragment), and SREP-PAT-OV (the algorithm with Step 5, where the text fragments from which relations are extracted may overlap). The rationale of including 'OV' variants was to estimate the number of potentially missed extractions by the base versions of the algorithm.

3.1 Test Corpus

In order to create the Test Corpus 238 sentences (either first or second sentence) from on-line news articles in Polish published during May 2015 were randomly selected using the *Europe Media Monitor*. These sentences cover various domains, including economy, finances, world and local politics, sports, culture and crisis situations. The main motivation behind the selection of initial sentences was due to our particular interest in the extraction of relations related to the main events of the news articles (Tanev et al., 2008). Figure 1 shows the histogram for sentences length in the test corpus. Nearly 50% of the sentences consists of 15 or more tokens which reflects the complexity level.

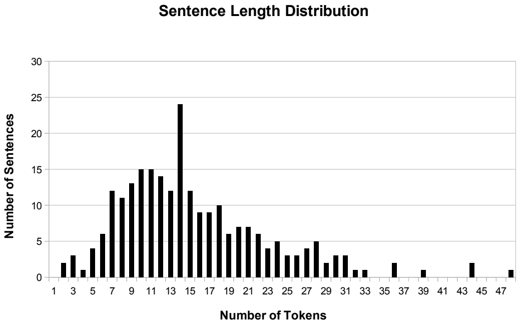

Figure 1: Sentence length distribution.

For each sentence in the test corpus 'to-be-extracted' relation triples were manually created. This task was accomplished by one human annotator. In total 616 relation triples were annotated, i.e., on average there are more than two re-

[4] Analogously to Step 2, the patterns for Step 5 were created via identification of the most prevalent constructions in the test corpus.

[5] The word '*to*' is a pronoun (meaning either 'it' or 'this') that can be used to express 'is-a' relation in Polish.

lations per sentence. It is important to note that n-nary (where $n > 3$) relations (e.g., *X* took place in *Y* at *Z*) were annotated as $n - 1$ triples accordingly: (*X*,*took place in*, *Y*) and (*X*,*took place at*,*Z*). For instance, for the sentence '*Lechia Gdańsk okazała się znacznie lepsza od APOEL-u Nikozja i pokonała go w meczu sparingowym aż 4:1 (1:0)*' (Lechia Gdańsk turned out to be significantly better than APOEL Nicosia and defeated it in a friendly match 4:1 (1:0)) in the test corpus the following two annotations[6] are made:

```
(Lechia Gdańsk,okazała się lepsza od,
 APOEL-u Nikozja)
```

```
(Lechia Gdańsk,pokonała,APOEL-u Nikozja)
```

The system does not lemmatize the arguments, i.e., the arguments in the returned triples are 1:1 copy of the surface forms in the text, e.g., *APOEL-u Nikozja* instead of *APOEL Nikozja*.

3.2 Experiments

Figure 2 shows the precision-recall curves for the *exact relation extraction task*[7] for the four configurations: SREP, SREP-PAT, SREP-OV and SREP-PAT-OV (see 2), computed by varying the confidence threshold. Somewhat unsurprisingly, one can observe that the overall results for exact matching are rather poor, in particular as regards recall. The version of the algorithm that includes the application of generic lexico-syntactic patterns (SREP-PAT) performs better than the version without (SREP) in terms of both recall and precision. Furthermore, one can observe a small boost in recall (at the cost of lowering precision figures) when 'overlapping' was allowed (SREP-PAT-OV), which indicates an area where improvement could be made.

In order to have a more in-depth picture of the error types we have computed precision-recall curves for the subtask of exact relation extraction task, namely, the *relation phrase extraction task*, which are depicted in Figure 3. One can observe significant improvement as regards both precision and recall vs. extracting entire relations, in particular for SREP and SREP-PAT configurations, for which the figures still lag behind the ones reported for relation phrase extraction for English (Fader et al., 2011) but are getting closer.

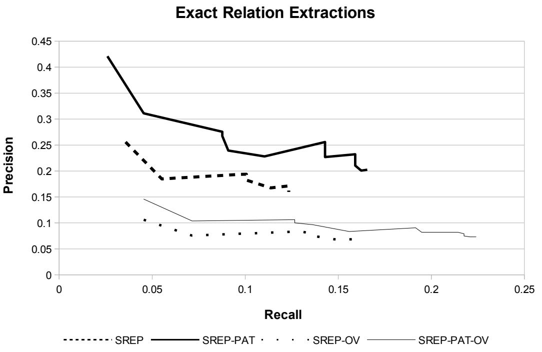

Figure 2: Precision-Recall curves for the exact relation extraction task.

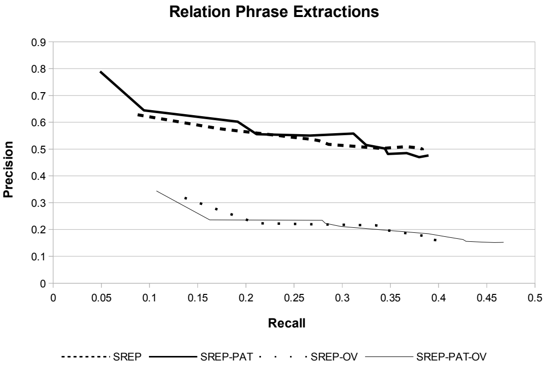

Figure 3: Precision-Recall curves for the relation phrase extraction task.

One can also conclude from Figure 3 that a significant number of errors stems from non correct extraction of the arguments of a relation. To study the problem more thoroughly we also computed the precision-recall curves for the *fuzzy relation extraction task*, in which an extracted triple (*X*,*rel*,*Y*) is considered to be correct if *rel* is identical with the corresponding value in the test corpus, whereas *X* and *Y* are similar to the corresponding values in the test corpus, i.e., the string distance between the extracted values and the correct ones in the test corpus is relatively small. For the purpose of computing string distance we used the *longest common substrings* distance metric (Navarro, 2001). Figure 4 presents the precision-recall curves for the fuzzy extraction task, where SREP-PAT-FUZZY-2 curve corresponds to a variant of fuzzy matching, in which relation phrase may also slightly differ from the

[6] One corresponding to 'being better' and one to 'winning a match' relation.

[7] Relation phrase and both arguments have to be identical with the corresponding annotation in the test corpus.

relation phrase in the test corpus. Although both precision and recall figures are higher vs. figures for exact relation matching, there is an indication (cf. Figure 3) that there is still a fraction of extracted relations for which the extraction of at least one of the arguments entirely failed, i.e., the error is not related to mismatching left/right boundary of the NP representing the argument.

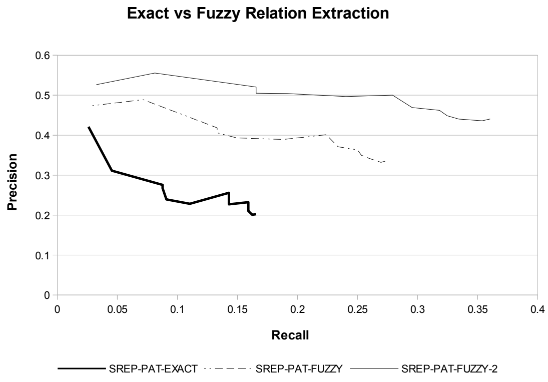

Figure 4: Fuzzy vs. exact relation extraction.

The analysis of the errors for the SPAT-PAT-FUZZY-2 configuration (i.e., errors that go beyond simple mismatching of the left/right boundaries of the text pieces that are to be extracted (both arguments and relation phrases)), revealed that: 34.8% of the errors are related to the extraction of triples that do not represent relations at all; 23.3% of the errors are due to the failure of extracting the first argument correctly (subject of the predicate); 14.0% of the errors are due to extracting *arg1* as *arg2* and vice versa; 7.0% of errors constitute errors, in which *arg2* is wrongly extracted; whereas remaining errors cover issues related to more significant mismatch of left/right margin of *arg1*, *arg2* or the relation phrase itself.

The main cause of missed relations was due to, i.a.,: (a) relation phrase not being present in the text between arguments (36.7%); (b) non-contiguous relation phrase structure (28.3%)[8]; (c) non-matching of POS-based patterns for detection of relation phrases (10.4%); and (d) non handling of constructions, in which arguments of the relations are "embraced" in verbs (8.9%)[9].

4 Conclusions and Outlook

We presented initial experiments on developing a linguistically-lightweight tool for open relation extraction for Polish that is an adaptation of an existing approach to open relation extraction for English. An evaluation carried out on a small set of sentences randomly extracted from Polish online news and a coarse-grained error analysis revealed that: (a) precision/recall figures for relation phrase extraction are promising, although a significant part of errors is due to extracting triples that do not represent any relations (ca. 35%), (b) performance of the extraction of relation arguments needs to be significantly improved as this is the main cause of errors, although, the observed errors did not result only from incorrect NP boundary detection[10], but also due to errors of different nature, e.g., extracting *arg1* as *arg2* and vice versa (14%).

We believe that the work in progress reported in this paper constitutes useful source of knowledge for researchers aiming at working on OIE for Slavic languages. In particular, the linguistically-poor approach to open relation extraction and the accompanying performance figures presented here could serve as a baseline to use against which to compare more sophisticated solutions.

Apart from improving the overall approach and fine-tuning the underlying resources, future work could possibly encompass integration of a mechanism to: (a) aid detecting argument boundaries, e.g., as the one in (Etzioni et al., 2011) and (b) decompose sentence into parts that belong together (Bast and Haussmann, 2013), but without deploying linguistically sophisticated tools, e.g., dependency parsers, in case one is interested in developing a Web-scale solution. Most likely, some of the identified problems could be tackled through the deployment of additional linguistic processing modules for Polish, e.g., a named-entity recognition component (Savary and Piskorski, 2011) could be used to improve NP boundary detection, while deployment of even a rudimentary co-reference resolution mechanism (Broda et al., 2012) could potentially help to handle zero anaphora to increase the recall. Finally, instead of relying on full-form lexica for computing POS information, full-fledged POS taggers could be deployed (Piasecki, 2007; Acedański, 2010; Radziszewski, 2013).

[8] Although some of the patterns in Step 5 of the algorithm do cover such cases.

[9] Polish is a null-subject language. No mechanism for detecting null-subjects was used.

[10] It constitutes one of the core problems while developing IE solutions for Polish.

Acknowledgments

We are indebted to Hristo Tanev from the Joint Research Centre of the European Commission, who provided the corpus of news articles in Polish.

References

Szymon Acedański. 2010. A morphosyntactic Brill tagger for inflectional languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *IceTAL*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14.

Douglas E. Appelt. 1999. Introduction to information extraction. *AI Commun.*, 12(3):161–172.

Hannah Bast and Elmar Haussmann. 2013. Open information extraction via contextual sentence decomposition. In *Proceedings of ICSC'13*, pages 154–159.

Bartosz Broda, Łukasz Burdka, and Marek Maziarz. 2012. Ikar: An improved kit for anaphora resolution for Polish. In Martin Kay and Christian Boitet, editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers*, pages 25–32.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 355–366.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. 2011. Open information extraction: The second generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 3–10.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, ROBUS-UNSUP '12, pages 10–18.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 523–534.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

Maciej Piasecki. 2007. Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167.

Adam Przepiórkowski. 2007. Slavonic information extraction and partial parsing. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL 2007, pages 1–10.

Adam Radziszewski. 2013. A tiered CRF tagger for Polish. In H. Rybiński M. Kryszkiewicz M. Niezgódka R. Bembenik, Ł. Skonieczny, editor, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.

Agata Savary and Jakub Piskorski. 2011. Language resources for named entity annotation in the national corpus of Polish. *Control and Cybernetics*, 40(2):361–391.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *Proceedings of the 13th International Conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems*, NLDB '08, pages 207–218.

Marcin Woliński, Marcin Miłkowski, Maciej Ogrodniczuk, and Adam Przepiórkowski. 2012. Polimorf: a (not so) new open morphological dictionary for Polish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 860–864, Istanbul, Turkey.

Alina Wróblewska and Marcin Sydow. 2012. DEBORA: dependency-based method for extracting entity-relationship triples from open-domain texts in Polish. In *Foundations of Intelligent Systems - 20th International Synposium (ISMIS) 2012, Macau, China, December 4-7, 2012, Proceedings*, pages 155–161.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127.

Clarissa Castellã Xavier, Vera Lúcia Strube de Lima, and Marlo Souza. 2015. Open information extraction based on lexical semantics. *Journal of the Brazilian Computer Society*, 21(4).