Extraction of *part-whole* relations from Polish texts based on Wikipedia and Cyc

Aleksander Smywiński-Pohl[1,2]

[1]Department of Management and Social Communication
Jagiellonian University
[2]Department of Computer Science, Electronics and Telecommunication
AGH University of Science and Technology

Abstract

The paper describes an algorithm used to extract instances of semantic relations from text. It utilizes Wikipedia and Cyc as its primary resources. Its aim is semi-automatic construction of extraction patterns, that cover both formal and semantic features. This algorithm was applied to the problem of extraction of *part-whole* relationship from Polish texts. The reported precision reaches 80-90% while coverage is in the range of 50-60%. These are one of the first results reported for the specified problem and also they are one of the very few results regarding application of Cyc in the problem of relation extraction.

1. Introduction

Information Extraction (IE) is the subfield of Natural Language Processing (NLP) concerned with acquiring structured data from textual resources. The methods of IE are usually lightweight, compared to fully-fledged NLP, based on deep parsing and sophisticated semantic theories. Yet its results can be useful, especially when compared to uniformed Information Retrieval (IR). While in IR the user is presented with a set of results matching her query, in IE the results have a compact form suitable for storage in knowledge sources such as relational databases.

Relation extraction (RE) is a task defined within IE, concerned with the identification of selected semantic relations, such as membership or parenthood. It is a crucial task, since useful information usually has the form of relations between objects. A hypothetical RE system could read newspapers in order to store *country – head of government* relationships in a relational database.

This paper presents an algorithm for relation extraction operating within the field of *closed* IE. The user has to indicate which relation he or she is interested in and then the algorithm tries to learn features (both formal and semantic) useful for recognizing that relation. The key features of the algorithm are as follows: the extracted expressions are grounded in a semantic dictionary based on Wikipedia, the semantic types of the expressions are defined using the terms of the Cyc ontology (Lenat, 1995), the manual labour of the user is reduced and simplified to yes/no questions regarding the presence of the extracted relation, the algorithm uses subsumption relation when deciding if the semantic constraints are satisfied and it was tested on samples of Polish texts. The presented experiments were carried for the *part-whole* relations, but the algorithm as well as the employed knowledge sources are general enough, allowing it to be extended for the other semantic relations.

2. Resources

We use two primary resources to accomplish the task of relation extraction: Polish Wikipedia and Cyc. We also utilize the English Wikipedia as a bridge between these two resources. We use Polish corpus from the Polish Academy of Sciences (IPI PAN corpus) (Przepiórkowski, 2006) to find examples of expressions connected by the relation and a corpus of short news from the Polish News Agency (PAP) as an open-domain testing corpus.

The Polish Wikipedia is a basis for a semantic dictionary defining a large number of general concepts as well as proper names. The second group of expressions is particularly important from the point of view of closed IE: in *instance-oriented* IE proper names are the primary means for referring to individual objects. Thus, semantic dictionaries like WordNet (Fellbaum, 1998), with their limited coverage of proper names, are not enough to ground meaning of the extracted expressions. Moreover, thanks to the development of Semantic Web, with the central role of DBpedia (a Wikipedia-derived knowledge base) (Auer et al., 2007) it is now pretty easy to share the meaning of the extracted data, assuming that they are grounded in DBpedia.

The role of Cyc is manifold – first of all the Cyc ontology is used to classify the extracted expressions as well as ground the meaning of the extracted relations. This allows leveraging the vast body of knowledge encoded in this ontology (like taxonomical relations for concepts *and* for relations). Cyc is also a source of examples of pairs of concepts connected by the relation that has to be extracted. Its taxonomical structure is leveraged in two steps of the algorithm. Firstly, when the examples are searched for in the corpus, the set of concept pairs provided for given relation is automatically expanded with their specializations. Secondly, when the semantic constraints imposed on the arguments are verified, allowing for more specific types to be assigned to the extracted expressions.

Both of these resources are used together thanks to the fact that the articles in the Polish Wikipedia were classified into the conceptual schema of Cyc. This was achieved in two steps. Firstly, the contents of the English Wikipedia was automatically classified using Cyc terms as semantic categories of the articles (Pohl, 2012a). Then thanks to the presence of interlingual links it was possible to cast the classification from the English to the Polish Wikipedia. To improve the coverage of the classification two heuristics were implemented. In the first *infoboxes* from the Polish Wikipedia were manually mapped to the Cyc terms (following the classification method implemented in DBpedia). In the second heuristic the articles included in *Urodzeni w N* (Eng. *N births*) and *Zmarli w N* (Eng. *N deaths*) categories were classified as describing people. As a result 80% of the articles were classified.

3. Pattern building

The semantic relation extraction is based on the construction of extraction patterns. These patterns are built semi-automatically (the process is depicted on Figure 1):

1. At the beginning the user selects a relation **r** that will be extracted from the text. The primary requirement is that the relation is defined in Cyc and it has a non-empty extension (pairs of symbols connected by that relation).

2. In the next step the relation extension is exported from Cyc. It is a set of symbol pairs ($\mathbf{s}_a$,$\mathbf{s}_b$) representing examples of the relation.

3. The exemplary sentences containing the references to the symbols are searched for in the IPI PAN corpus. The search has two stages, each devoted to one of the symbols in the pair. This allows for matching direct occurrence of the symbol or one of its specializations. A set of examples $\mathbf{z}_{PAN}$ is constructed.

4. This set is filtered in order to remove incorrect, e.g. duplicated examples. A cleaned set $\mathbf{z'}_{PAN}$ is constructed.

5. Based on that set the *formal patterns* – **fp** – are constructed. They cover morphological and syntactic features, such as the grammatical category of the expression, the order of the arguments, the words that occur between them, etc.

6. Identical patterns are merged preserving the number of distinct examples and distinct expressions that occurred in the examples, forming the set of *generalized formal patterns* – **gfp**.

7. A target corpus (PAP) is morphosyntactically tagged using Concraft tagger (Waszczuk, 2012) and semantically disambiguated using an algorithm described in (Pohl, 2012b). Sentences containing at least two recognized concepts are selected forming the $\mathbf{z}_{PAP}$ set.

8. The **gfp** patterns are matched against that corpus. As a result a set of tuples $\mathbf{z'}_{PAP}$ containing the natural language expressions corresponding to the arguments of the relation is identified.

9. A subset of the extracted tuples is reviewed by a human evaluator. The positive examples containing the relation are used to extract *semantic constraints* ($\mathbf{c}_a$,$\mathbf{c}_b$) of the relation. Since the expressions appearing in the examples may have many semantic categories attached, for each pair of categories in the Cartesian product of the classifications a separate pair of constraints is established.

10. The final extraction patterns – **ep** – are composed of the formal patterns and semantic constraints and are used to extract the instances of the relation from the text.

4. Pattern matching

The matching of the extraction patterns against the analyzed text resembles the steps in the pattern construction. The primary difference is the step that matches the expressions against the semantic constraints imposed on the arguments. The patterns are matched as follows:

1. In the first step the text is tagged with Concraft tagger $\rightarrow \mathbf{T}_{tag}$.

2. Then the text is sematically disambiguated using the same algorithm as in step 7 $\rightarrow \mathbf{T}_{sem}$.

3. The formal patterns are matched against the disambiguated expressions, identifying a set of sentences that potentially include the semantic relations $\rightarrow \mathbf{Z}$.

4. The candidate pairs of expressions are checked against the semantic constraints – either strictly (the semantic categories of the arguments have to be the same as the constraints) or using subsumption (the semantic categories of arguments have to generalize to the constraints).

The result of the extraction is a set of pairs of expressions, together with offsets and span of their occurrences, target Wikipedia articles, their semantic types and an assignment of the expressions to the formal arguments of the relation.

5. Experiments with *part-whole* relationship

We have conducted experiments with the *part-whole* relationship. Although there is `#$parts` predicate in Cyc that correspond directly to the general notion of *part-whole* relation, we have used `#$anatomicalParts` predicate as its paradigmatic example. It is well defined in Cyc and has a number of examples (94) expressed via `#$relationAllExists` meta-predicate[1]. For instance `#$Bear` is connected with `#$Claw`, `#$Bull-Cattle` with `#$Horn-AnimalBodyPart`, `#$Cactus` with `#$Thorn`, etc. These concepts were already translated into Polish in our earlier research[2].

In the next stage we have searched for exemplary sentences containing these concepts in the IPI PAN corpus. In order to generate a larger number of examples, we have used not only the concepts present in the assertions, but also their specializations. 4 scenarios were implemented varying on the fact, whether the concepts should appear directly in the sentence or they should be represented by their children (i.e. more specific concepts).

The results in Table 1 show how many examples were generated when only one of the concepts had to appear in the sentence (i.e. they show the characteristic of the IPI

[1] We have used ResearchCyc since OpenCyc does not contain such assertions.

[2] `https://github.com/apohllo/polish-cyc`

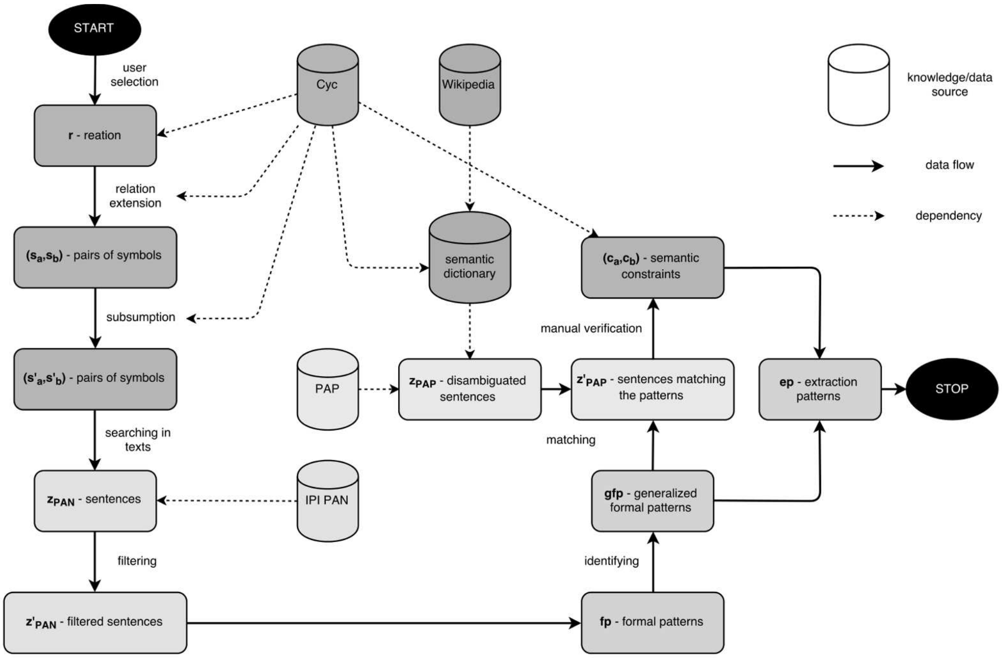

Figure 1: A schema showing the construction of extraction patterns.

Table 1: The number of examples generated in the first stage of example generation.

Argument	**Match**	**Count**
whole	direct	73416
part	direct	100443
whole	child	446828
part	child	27641
Total		**648328**

PAN corpus w.r.t. the anatomical domain). The results given in Table 2 show the second stage of the example selection. Here the *First argument* and the other indices refer to the order in which the arguments were searched for in the corpus, not their position within the relation. The number of examples is much smaller than in Table 1, showing that the example generation method is pretty selective. It also shows, that various implemented searching scenarios largely increased the number of examples (from 695 for the direct – direct method up to 10218 for the combination of all methods).

In the next step the examples were filtered in order to remove duplicated sentences (there is no notion of sentence or document id in the IPI PAN corpus) and matches where the expressions were parts of different sentences. As a result 3054 unique examples were preserved from the initial set of 10218 examples.

The next step was the construction of the formal patterns of the semantic relation as appearing in the collected examples. The patterns included the grammatical classes of the arguments, the values of gender, number and case[3], the order of the arguments[4] and (possibly empty) inner context appearing between the arguments.

Table 2: The number of examples generated in the second stage of example generation.

First argument	**Match 1**	**Match 2**	**Count**
whole	direct	direct	130
part	direct	direct	565
whole	direct	child	16
part	direct	child	6074
whole	child	direct	249
part	child	direct	61
whole	child	child	37
part	child	child	3086
Total			**10218**

E.g. for the expressions:

> W „Akwarium” można dostrzec fragmenty *skóry* ***makreli***... (It is possible to see fragments of ***mackerel's*** *skin* in the „aquarium”...)[5],

the pattern described in Table 3 was constructed. It means that the *whole* was to the right of the *part*, the *whole* was a singular (sg), feminine (f) noun (subst) in dative (dat), the inner context was empty, and the *part* was also a singular, feminine noun in dative. It should be noted that some of the expressions spanned many words. In such cases the tags were taken from the head of the expression equated with the first noun counting from the left.

[3] They were left empty if they were not applicable for the grammatical class.

[4] Whether *whole* was to the left or to the right of the *part*.

[5] The *whole* is marked as bold italic, while the *part* as italic.

Table 3: An example formal pattern constructed from one of the examples appearing in text.

Feature(s)	Value
order	`right_left`
whole features	`subst:dat:sg:f`
inner context	`--`
part features	`subst:dat:sg:f`

Since many of the examples included the same patterns, it was possible to identify them and to collect usage counts of the patterns. We have used two types of counts – the total number of examples including given pattern and the number of distinct argument values matching given pattern. It turned out that the second count was a better indicator of the pattern quality. The reason for that was caused by fact, that genuinely different evidences were used to build the pattern. In the first case the count reflected only the overall popularity of the pair of expressions, which might have been connected by different relation than meronymy (e.g. like in proverbs or figurative uses). As a result 2319 unique patterns were identified in the corpus, but only 156 were preserved, since we have set the minimum requirement for distinct argument values to 2.

The 156 generalized formal patterns were matched agains the PAP corpus. The matching was preceded by morphosyntactic tagging and semantic disambiguation. As a result the formal patterns were not matched against raw stream of tokens with attached tags, but only those expressions that were disambiguated. This trick greatly reduced the amount of computation required for pattern matching and allowed matching the patterns not only against words, but also multi-word units (the head was determined in the same way as in the pattern construction). The matching of the patterns against 50 thousand short news brought almost 21 thousand matches, mening that in more that 40% news a potential relation was detected.

To extract the semantic constraints from examples, it was required that the expressions were supplemented with semantic categories. Since the construction of the semantic dictionary was automatic, the categories were defined only for 80% of the entries. Moreover, in the case of pairs of entries, the probability for both of them to have a semantic category assigned was even lower. It turned out that only in 13 thousand of the examples (∼62%) the categories were defined for both of the arguments. Moreover, to improve the quality of the semantic disambiguation of the entries, a minum confidence threshold (0.2) was set for the results, reducing the number of available examples to 6816. 10% of these examples were reviewed manually, to check if the *part-whole* relationship was present. The 116 positive examples (17%) were used to define the semantic constraints of the relation, totaling in 405 pairs of constraints (caused by the fact that many of the expressions have had multiple semantic categories assigned). For instance (`#$IndependentCountry`, `#$IntelligenceAgency`) was defined as a pair of semantic constraints of the relation.

6. Results

As it is stated in the previous paragraph, only 17% of the examples matched against the formal patterns included an occurrence of the *part-whole* relationship. This primary result clearly indicates that the generalized formal patterns were not enough to capture the features of this semantic relation. Thus, the requirement for the semantic constraints was amplified. In the second part of the experiment, we wanted to find out if the constraints defined on the basis of a small sample were selective enough to properly identify the relation. Also we wanted to check if there was any difference in the performance of the algorithm between a scenario when the constraints were matched directly or with the help of the subsumption relation.

The results of the second part of the experiment are summarized in table 4. Precision (Pr), recall (Rc) and F_1 score are defined as in IR. The differences in the results stem from two parameters of the algorithm:

- **Gen.** – indicates if the sumbsumption relation was used to satisfy the constraints (+ means that it was used),
- **Id.** – indicates if the pair was not extracted if the semantic categories of both arguments were identical (+ means that such pairs were not extracted).

The second parameter was included as an outcome of manual inspection of the extraction results, where appositives, such as *deputy director John Smith*, were causing many invalid extractions. Thus, a heuristic forbidding extraction when both of the extracted expressions have at least one semantic category in common was implemented.

Table 4: The results of the extraction of *part-whole* relation from PAP corpus based on extraction patterns.

Gen.	Id.	Pr [%]	Rc [%]	F_1 [%]
–	–	52,5	51,5	52,0
–	+	**89,0**	49,9	64,0
+	–	49,0	**61,5**	54,5
+	+	78,6	57,6	**66,5**

It turns out that the best result (in terms of F_1 score) was achieved for the variant of the algorithm with both parameters turned on. This means specifically that the coverage gains from including subsumption relation when matching the semantic constraints outperformed the precision drops. On the other hand the difference between similar configuration, when the subsumption was not allowed, was small (2,5 percentage points), while the precision reached almost 90%. Thus, including subsumption when matching constraints, if the precision is the primary concern, is not the best idea. On the other hand the results for the second parameter regarding the identical semantic categories clearly show, that this simple heuristic was very important, since in both cases the precision was improved by more than 29 percentage points.

7. Related work

Although the literature on the problem of relation extraction is vast, we will discuss only two similar efforts: the work of Girju and Badulescu (Girju et al., 2006) and Piasecki, Szpakowicz and Broda (Piasecki et al., 2009), since we find them to be most relevant for the presented algorithm.

Girju has extracted the *part-whole* relationship from English texts. The algorithm was based on the English WordNet and was not much concerned with the proper names. The reported precision reached 83% while the coverage ranged from 72% to 98%.

Piasecki et al. have implemented several algorithms for extracting hypernymy (subsumption) relations in the context of automatic building of the Polish WordNet. The first one was based on manually constructed patterns (following (Hearst, 1992)). Its precision ranged from 10% to 74% depending on the number of evidences found in the corpus, but the reduction in the number of detected relations for the most precise results was more than 40-fold. The second implemented method was based on seed examples (following (Pantel and Pennacchiotti, 2006)) and was most similar to the method presented in this paper. The achieved precision ranged from 39% to 59%, but the coverage was not reported.

Although the presented efforts were the most similar to our algorithm, it is rather hard to directly compare the results. When comparing with (Girju et al., 2006) we have to take into account the differences in English (positional language) and Polish (inflectional language), which have great impact on the problem of relation extraction. On the other hand the method described in (Piasecki et al., 2009), have two features that make them distinct from our case: a different relation was extracted (hypernymy) and the results were evaluated in the context of ontology building rather than semantic mark-up.

8. Conclusions

The overall conclusions from the presented experiments are as follows. First of all the application of semantic constraints in the problem of *part-whole* relationship extraction has great impact on the precision of the extraction, changing it from 17% (when only formal patterns were used in the extraction) up to 90% (for the most precise variant of the algorithm). The second conclusion is that it is possible to extract the relations with relatively high precision (80-90%), when a simple heuristic checking the equality of semantic categories of the arguments is in operation. In future we will test the performance of that approach in other language (especially English), we will work on removing the requirement for manual intervention in the construction of the extraction patterns and test its performance on other semantic relations.

The results of this research can also be implemented in a system for information retrieval from audio-video signal that is constructed with the participation of AGH. The speech in the signal is automatically recognized and transcribed, allowing for application of IE techniques, such as relation extraction.

9. Acknowledgment

This work was supported by DOB-BIO6/22/133/2014 grant from the Polish National Center for Research and Development.

10. References

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives, 2007. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux (eds.), *The Semantic Web*, volume 4825. Berlin, Heidelberg: Springer, pages 722–735.

Fellbaum, Christiane, 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Girju, Roxana, Adriana Badulescu, and Dan Moldovan, 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Hearst, Marti A, 1992. Automatic acquisition of hyponyms from large text corpora. In Antonio Zampolli (ed.), *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics.

Lenat, Douglas B., 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Pantel, Patrick and Marco Pennacchiotti, 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In Marine Carpuat and Kevin Duh (eds.), *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Piasecki, Maciej, Stanisława Szpakowicz, and Bartosz Broda, 2009. *A WordNet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Pohl, Aleksander, 2012a. Classifying the Wikipedia Articles into the OpenCyc Taxonomy. In Giuseppe Rizzo, Pablo Mendes, Eric Charton, Sebastian Hellmann, and Aditya Kalyanpur (eds.), *Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*.

Pohl, Aleksander, 2012b. Improving the Wikipedia Miner Word Sense Disambiguation Algorithm. In Maria Ganzha and Marcin Paprzycki (eds.), *Proceedings of Federated Conference on Computer Science and Information Systems 2012*. IEEE.

Przepiórkowski, A., 2006. The potential of the IPI PAN Corpus. *Poznan Studies in Contemporary Linguistics*, 41:31–48.

Waszczuk, Jakub, 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In Martin Kay and Christian Boitet (eds.), *Proceedings of COLING*.