Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources

Darja Fišer, Senja Pollak, Špela Vintar

University of Ljubljana, Department of Translation
Aškerčeva 2, SI – 1000 Ljubljana
E-mail: {darja.fiser, spela.vintar}@guest.arnes.si, senja.pollak@ff.uni-lj.si

Abstract

The paper presents an innovative approach to extract Slovene definition candidates from domain-specific corpora using morphosyntactic patterns, automatic terminology recognition and semantic tagging with wordnet senses. First, a classification model was trained on examples from Slovene Wikipedia which was then used to find well-formed definitions among the extracted candidates. The results of the experiment are encouraging, with accuracy ranging from 67% to 71%. The paper also addresses some drawbacks of the approach and suggests ways to overcome them in future work.

1. Introduction

Extracting domain-specific knowledge from texts has become one of the most proliferate areas of natural language processing and involves tasks such as terminology extraction, extraction of semantic relations, gloss or definition extraction, named entity recognition and other tasks aimed at harvesting meaningful items of knowledge. Especially the construction of semantic resources to be used in knowledge applications requires an insight into text that goes beyond traditional levels of automated analysis.

In this paper we present a method of extracting definition candidates from domain-specific corpora using a variety of approaches, including morphosyntactic patterns, terminology recognition and semantic tagging. In order to distinguish between positive and negative examples of definitions we first compiled a training set of definitions and non-definitions from Slovene Wikipedia and used it to build a classification model, then classified the candidates and evaluated the model. The approach thus exploits a structured data source, i.e. Wikipedia with its standard article format, to learn rules which help us extract knowledge from an unstructured resource such as a domain-specific corpus.

A definition is a passage describing the meaning of a term; a word, phrase or other set of symbols (Wikipedia). The traditional Aristotelian notion of a definition further specifies its structure as *per genus et differentiam*, meaning that the term is to be defined by its hypernym or broader term (genus) and the set of characteristics that distinguish the term from similar concepts. This structure is implied also in the pattern *[NP] is_a [NP] which [...]* often used in definition extraction. Definition extraction is a well-researched topic in natural language processing and is closely related to the extraction of semantic relations, with the latter being a broader field of research and a particularly crucial part of automated ontology construction. While both are concerned with identifying a set of related concepts in text, definitions are structurally more formal and usually consist of a single sentence, whereas semantic relations can span over much broader text segments or indeed over the entire text.

There are two main approaches to definition extraction from large specialized corpora and web resources; the rule-based approaches and machine learning approaches, whereby recent studies often combine both. Rule-based approaches are based on pattern matching using mainly syntactic and lexical features, but also paralinguistic and/or layout information. Hearst (1992) proposed a method for extraction of hyponym relations from large corpora, based on a set of lexico-syntactic patterns. Along similar lines other types of relations (synonym and hyperonym) have been addressed (eg. Malaisé et al. 2004). Pattern-based methods are still important in the field of automatic definition extraction (e.g. Muresan and Klavans 2002; Walter and Pinkal 2006; Storrer and Wellinghoff 2006; Del Gaudio and Branco 2007). To define the matching patterns automatically, bootstrapping techniques can be used (e.g. Riloff & Jones 1999; Walter 2008). While definitions in structured resources such as dictionaries and encyclopaedias usually comply with the formal definition template, many candidate definitions extracted from running texts show more variety in the linguistic structures they contain. To overcome this drawback of the pattern-based approach we complement it with semantic annotation and term extraction to identify definition candidates.

The second line of related work is based on machine learning (ML), understanding the definition extraction as a classification task. Machine learning techniques are often used in combination with pattern recognition approaches. Experiments show that ML can help determine which definition candidates are relevant and well-formed, using standard classifiers such as Naive Bayes, Decision Trees, Support Vector Machines (Chang and Zheng 2007; Velardi et al. 2008; Fahmi and Bouma 2006), but also Balanced Random Forest (e.g. Kobilinski and Przepiorkowski 2008, Westerhout 2009) or genetic algorithms (Borg et al. 2009). Compared to rule-based approaches, ML techniques require more training data and have to deal with often unbalanced datasets.

Slovene is a morphologically rich language with free word order, and is spoken by a relatively small community. Nevertheless, much has been accomplished in the past few years in terms of NLP resources and tools. Extensive annotated corpora have been collected and there are efforts to provide lexico-semantic resources as well. For example, a wordnet for Slovene has been built to a large extent automatically, exploiting bilingual dictionaries (Erjavec and Fišer 2006), multilingual parallel corpora (Fišer 2007) and various web-based semantic resources (Fišer and Sagot 2008). In previous experiments we have successfully implemented term extraction as a way of automatically improving the domain coverage of Wordnet (Vintar and Fišer 2009) and to enrich it with multi-word expressions (Vintar and Fišer 2008).

This paper is structured as follows: in Sections 2 and 3 we lay out the procedure of learning definitions from Wikipedia and specialized corpora, incorporating semantic annotation with sloWNet, automatic term recognition and pattern matching. In Section 4 we describe the experiment settings as well as present the results which are evaluated on a manually annotated gold standard. The paper is concluded with a discussion of the results and plans for future work.

2. Learning Slovene definitions from Wikipedia

The purpose of our experiment is to extract well-formed definitions from a domain-specific corpus of Slovene texts, with the aim of integrating them into Slovene Wordnet. To learn the rules that will help us distinguish between well-formed and not-well-formed definitions we built a training set of positive and negative definition examples from Slovene Wikipedia, whereby we assume that the first sentence in the Wikipedia encyclopaedic article is the definition of the term. As negative examples we automatically selected sentences beginning with the target term from the remainder of the article. Thus, the sentence

Celica je strukturna in funkcionalna enota vseh živih organizmov.
[A cell is a structural and functional unit of all living organisms.]

represents a positive definition example, while the sentence

Celice so v povprečju velike 10-20 μm, s prostim očesom jih ne moremo videti.
[Cells are typically of sizes 10-20 μm and cannot be seen with the naked eye.]

is a negative definition example, because it begins with the target term but does not define it.

From the Slovene Wikipedia as of December 2009 which contained 162,500 articles we selected only well-formed pages and removed those that contained no text. Pages were morphosyntactically annotated and lemmatized with ToTaLe (Erjavec et al. 2005), then structurally parsed. From each article we selected the first sentence as the definition and another sentence containing the title term from the remainder of the page as the non-definition (see example above). We trained a learning algorithm on this training set containing 19,964 instances, whereby we experimented with decision trees and decision rules algorithms from the Weka data mining toolkit (Witten and Frank 2005). As features we used the most frequent definition and non-definition part-of-speech tags and lemmas.

3. Extracting definition candidates from specialized corpora

In order to extract definition candidates from unstructured resources we built a corpus of knowledge-rich texts from several natural science domains such as astronomy, physics, geography, botany etc. using the texts from the FidaPlus[1] reference corpus of Slovene. The specialized corpus contained primarily textbooks and popular science volumes targeting students and non-expert readers; the choice was made with a view on including as many defining contexts as possible.

Potential definitions were then extracted from the corpus on the basis of the following three hypotheses. We assume the sentence to contain a definition or have defining content if:

- The sentence starts with a sloWNet literal and contains at least one more literal from the same hypernymy chain (i.e. its hyponym or its hypernym);
- The sentence contains at least two domain-specific terms in the nominative case; and
- The sentence contains a defining morphosyntactic pattern *(NP[nominative] is_a NP [nominative].*

The hypotheses are intentionally relatively broad because we aim for good coverage and refrain from presupposing a single definition structure.

3.1 Extraction with Wordnet

The semantic annotator identifies words and phrases in the lemmatised corpus that are already included in the present version of sloWNet. In cases of nested terms we select the longer, i.e. more specific term. If a sentence is found to contain two sloWNet terms where one is the hypernym of the other and the sentence begins with one of the terms, we extract it as a definition candidate.

[1] http://www.fidaplus.net

In the example below we show semantic annotation only for the two related terms responsible for the extraction of the definition candidate. Annotations of other terms as well as annotations at PoS-level and lemmas were omitted for clarity.

<term id=ENG20-13313485-n>Diabetes</term> je <term id=ENG20-13268088-n>bolezen</term>, ki je posledica pomanjkanja inzulina, hormona, ki skrbi, da celice v telesu dobivajo glukozo (sladkor).
[Diabetes is a disease resulting from insulin deficiency, the hormone providing glucose (sugar) for body cells.]

3.2 Extraction using Automatic Term Recognition (ATR)

The term recognition module identifies potentially relevant terminological phrases on the basis of predefined morphosyntactic pattern (Noun + Noun[genitive]; Adjective + Noun etc.). These phrases are then filtered according to a weighting measure *W* which compares normalized relative frequencies of single words between the domain-specific corpus *D* and the reference corpus *R* (FidaPlus). The frequency of the candidate phrase is also part of the weighting measure:

$$W(a) = \frac{f_a^2}{n} \cdot \sum \left(\log \frac{f_{n,D}}{N_D} - \log \frac{f_{n,R}}{N_R} \right)$$

We do not remove named entities from the list of term candidates because acronyms and Latin names are an important part of the specialised vocabulary. If a sentence contains two or more single- or multi-word terms proposed by the ATR module, of which both are in the nominative case, we extract it as a potential definition.

<term score="80.45“>Ekvator</term> je najdaljši vzporednik, ki deli Zemljo na severno in <term score=”43.21”>južno poloblo</term>.
[The Equator is the largest circle of latitude dividing the Earth into the Northern and the Southern Hemispheres.]

3.3 Pattern-based definition extraction

Our last approach is the traditional one to definition extraction, although its main drawback is low recall especially if used on less structured texts. We used a single, relatively non-restrictive pattern NP[nominative] *je/so* NP [nominative], which apart from true definitions matches many non-defining general contexts. The definition candidates from all three settings were manually validated to enable the evaluation of the learning algorithm.

4. Results

To be able to classify the extracted definition candidates we first performed a series of learning experiments with different algorithms and feature parameters on the Wikipedia training set. Table 1 gives a summary of the results, whereby evaluation was performed by 10-fold cross-validation; we list the accuracy, precision, recall and F-measure of each setting.

The ORIG label refers to full part-of-speech tags as attributes, while MERGED means that PoS tags have been collapsed by deleting all irrelevant grammatical information and thus reducing the number of attributes. The _bin label refers to the binary (yes/no) values instead of numerical representation of the attribute frequency.

The best results are obtained either using original PoS in binary representation or merging PoS tags (in binary or AF representation). For the classifiers we decided to use the models built with the J48 decision tree, M=10, which on average performs best. The confusion matrix for the best run also shows that a large number of candidates are assigned the correct class (classification accuracy 82.7%). For definitions only we achieve 0.83 precision and 0.82 recall (F-measure 0.827).

The three definition extraction methods described above yielded over a thousand definition candidates in total which were then hand-validated. Manual validation showed that about a third of the extracted candidates were well-formed defnitions. Table 2 shows the number of candidates extracted with each method and its precision.

Manual validation of the candidates revealed the true complexity of the task. While Wikipedia definitions are mostly uniformly structured and correspond with the expected [NP] is_a [NP] form, the definition candidates extracted from the corpus display a much wider range of syntactic structures. Even more difficult was the distinction between definitions and non-definitions from the content perspective. In running text, new concepts are introduced and described in a varied and ever-changing manner. Sometimes a concept is defined by specifying what it is not, other times several concepts are defined or explained within the same context. This should be taken into account when interpreting the results of our method; in the manual evaluation many candidates were marked as non-definitions although they contained valuable knowledge about the concept and could be referred to as *defining contexts.*

SETS	**Instances**	**Attributes**	**NaiveBayes**	**J48 (default)**	**J48 (M=10)**	**JRIP (default)**	**PART (default)**
ORIG	19964	260	66.91% (0.691/0.669/0.659)	81.59% (0.816/0.816/0.816)	82.13% (0.821/0.821/0.821)	80.91% (0.891, 0.891, 0.891)	82.56% (0.825, 0.825, 0.825)
ORIG_bin	19964	260	73.85% (0.691/0.669/0.659)	82.38% (0.824/0.824/0.824)	82.2% (0.822/0.822/0.822)	80.6% (0.806, 0.806, 0.806)	81.88% (0.819, 0.819, 0.819)
MERGED	19964	188	62.64% (0.674/0.626/0.599)	82.51% (0.825/0.825/0.825)	**82.72% (0.827/0.827/0.827)**	81.68% (0.817/0.817/0.817)	82.72% (0.827, 0.827, 0.827)
MERGED _bin	19964	188	72.39% (0.724/0.724/0.724)	82.18% (0.82/0.82/0.82)	82.44% (0.824/0.824/0.824)	80.5% (0.805, 0.805, 0.805)	81.79% (0.818, 0.818, 0.818)

Table 1: Classification accuracy followed by (precision, recall and F-measure) on Wikipedia training set

	Def. candidates	True definitions	Precision
sloWNet	104	41	0.39
ATR	629	118	0.19
Patterns	311	98	0.31
Total/Av.	1044	257	0.29

Table 2: Definition candidates

	Patterns	ATR	SloWNet
MERGED + J48-M10	**69.45%** (0.701/0.695/0.697/0.709)	69.79% (0.698/0.698/0.698/0.673)	61.76% (0.603/0.618/0.6/ 0.717)
MERGED_bin + J48-M10	63.9% (0.648/0.64/ 0.643/ 0.651)	**71.06%** (0.706/0.711/0.708/ 0.667)	**66.67%** (0.66/ 0.667/0.65/ 0.636)
ORIG_bin + J48-M10	62.7% (0.648/0.627/0.635/ 0.65)	65.98% (0.664/0.66/ 0.662/ 0.651)	63.72% (0.625/0.637/0.617/ 0.59)

Table 3: Classification accuracy for the three test sets

Table 3 shows the results of the classification performed on the three test sets obtained through pattern-based extraction, automatic term recognition (ATR) and SloWNet annotation.

The accuracy of the classifier trained on Wikipedia definitions ranges from 62% to 71%, which roughly means that the algorithm correctly assigns the class to the majority of the instances. Simplifying part-of-speech tags by erasing irrelevant grammatical categories such as gender improves performance, whereas the effect of binary attribute values as opposed to absolute frequencies is less clear.

Since one of the aims of our experiment was to compare the three definition extraction methods, it is interesting to look at the precision of the classifier on definitions only. The highest score was achieved with the SloWNet annotated test set (P: 0.63 / R: 0.415 / F: 0.5), followed by patterns (P: 0.514 / R: 0.551 F: / 0.532) and ATR (P:0.46 / R:0.441 / F:0.452). Although these figures basically refer to the extent of compliance with the training set, the same ranking of the methods can be seen from manual validation (see Table 2). In other words, if a sentence contains a Wordnet term and its hypernym, and one of the two terms appears at the sentence initial position, it is much more likely for this sentence to be a true definition than a sentence which contains any two domain-specific terms at any position. On the other hand, the ATR method yields the most definition candidates and is likely to propose defining contexts which do not comply with the standard definition structure but may still be relevant for knowledge extraction.

5. Conclusions

In this paper we described an innovative and efficient approach to extracting definitions from unstructured domain-specific corpora using machine learning and a combination of other language processing methods. The basic hypothesis underlying the experiments described is that definitions are best learned from structured resources such as Wikipedia, and the knowledge gained in this way can then be exploited to mine definitions from larger unstructured resources.

We employ a semantically-rich approach using terminology extraction, semantic tagging with wordnet senses and pattern-based extraction as parallel methods of obtaining knowledge from corpora.

Results show that the approach yields numerous well-formed definitions that can be integrated into the Slovene wordnet or used for terminographic purposes. Among the drawbacks of our method is relatively low recall if we wish to retain a high precision, and the inherent interdisciplinarity of domain-specific corpora.

The experiment also revealed the fuzziness of the concept of definition itself, particularly when comparing encyclopaedic definitions with those found in running texts. Not only are the latter structurally more flexible, they are also register and context dependent. Thus, a plant-infecting parasite can either be defined according to its zoological taxonomy or according to its effects on the host, environment or man; in each of these cases the defining sentence may contain a different hypernym for the target term to be defined.

In our future work we plan to improve the learning algorithm by introducing other levels of information as features as well as by using active learning. To facilitate the evaluation and interpretation of the results we also need to establish clearer guidelines for what we consider to be definitions.

References

Borg, C., Rosner, M. and Pace, G. 2009. “Evolutionary algorithms for definition extraction.” In *Proceedings of the 1st International Workshop on Definition Extraction*, RANLP-09. 18 September 2009, Borovets, Bulgaria, pp. 26-32.

Chang, X. and Zheng, Q. 2007. “Offline definition extraction using machine learning for knowledge-oriented question answering.” In *Proceedings of the Third International Conference on Intelligent Computing,* ICIC'07, 21-24 August 2007, Qingdao, China. Communications in Computer and Information Science, Vol. 2, Springer Berlin Heidelberg, pp. 1286-1294.

Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. 2005. Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In *Proceedings of the 2nd Language & Technology Conference*, April 21-23, Poznan, Poland. , pp. 32-36.

Erjavec, T. and Fišer, D. 2006. “Building Slovene WordNet.” In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC’06, 22-28 May 2006, Genoa, Italy, pp. 1678-1683.

Fahmi, I. and Bouma, G. 2006. “Learning to identify definitions using syntactic features.” In *Proceedings of Workshop on Learning Structured Information in Natural Language Applications*, EACL'06, 3-7 April, Trento, Italy.

Fišer, D. 2007. “Leveraging parallel corpora and existing Wordnets for automatic construction of the Slovene Wordnet.” In *Proceedings of the 3rd Language and Technology Conference*, LTC'07, 5-7 October 2007, Poznan, Poland, pp. 162-166.

Fišer, D. and Sagot, B. 2008. “Combining multiple resources to build reliable Wordnets”. In *Proceedings of the 11th International Conference on Text, Speech and Dialogue*, TSD’08, 8-12 September 2008, Brno, Czech Republic, pp. 61-68.

Gaudio, D. and Branco, A. 2007. Automatic extraction of definitions in Portuguese: A rule-based approach. *Progress in Artificial* Intelligence. *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, pp. 659-670.

Gaudio, R. D. and Branco, A. 2009. Language independent system for definition extraction: first results using learning algorithms. In *Proceedings of the 1st International Workshop on Definition Extraction*, RANLP-09. 18h September 2009, Borovets, Bulgaria, pp. 33-39.

Hearst, M.A. 1992. “Automatic acquisition of hyponyms from large text corpora.” In *Proceedings of the 14th International Conference on Computational Linguistics*, COLING'92, 23-28 July 1992, Nantes, France, pp. 539-545.

Kobyliński, L. and Przepiórkowski, A. 2008. “Definition extraction with balanced random forests.” In *Proceedings of the 6th International Conference on Natural Language Processing*, GoTAL 2008, Springer Verlag, pp. 237-247.

Malaisé, V., Zweigenbaum, P. and Bachimont, B. 2004. “Detecting semantic relations between terms and definitions.” In *Proceedings of the 3rd International Workshop on Computational Terminology*, CompuTerm'04, COLING'04, 29 August 2004, Geneva, Switzerland, pp. 55-62.

Muresan, S. and Klavans, J. 2002. A method for automatically building and evaluating dictionary resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, LREC'02, 29-31 May 2002, Las Palmas, Spain.

Riloff, E. and Jones, R. 1999. Learning dictionaries for information extraction using multi-level boot-strapping. In *Proceedings of The Sixteenth National Conference on Artificial Intelligence*, AAAI-99, 18–22 July 1999, Orlando, Florida. pp. 474-479.

Storrer, A. and Wellinghoff, S. 2006. Automated detection and annotation of term definitions in German text corpora. In *Procedings the 5th International Conference on Language Resources and Evaluation*, LREC’06, 22-28 May 2006, Genoa, Italy.

Velardi, P., Navigli, R. and D'Amadio, P. 2008. “Mining the Web to Create Specialized Glossaries”. *IEEE Intelligent Systems*, (23/5), IEEE Press, pp. 18-25.

Vintar, Š. 2004. “Comparative Evaluation of C-value in the Treatment of Nested Terms.” In *Memura 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications* (LREC 2004), 54--57.

Vintar, Š. and Fišer, D. 2008. “Harvesting Multi-Word Expressions from Parallel Corpora.” In *Proceedings of the 6th International Conference on Language Resources and Evaluation,* LREC'08, 28-30 May 2008, Marrakech, Morocco.

Vintar, Š. and Fišer, D. 2009. “Adding Multi-Word Expressions to sloWNet.” In *Proceedings of the 5th MONDILEX Workshop on Research Infrastructure for Digital Lexicography*, 14 October 2009, Ljubljana, Slovenia.

Walter, S. and Pinkal, M. 2006. Automatic extraction of definitions from German court decisions. In *Proceeding of the ACL'06 Workshop on Information Extraction beyond the Document,* 22 July 2006, Sydney, Australia, pp.20-26.

Walter, S. 2008. “Linguistic description and automatic extraction of definitions from German court decisions.” In *Proceedings of the 6th International Conference on Language Resources and Evaluation,* LREC'08, 28-30 May 2008, Marrakech, Morocco.

Westerhout, W. 2009. “Definition extraction using linguistic and structural features.” In *Proceedings of the 1st International Workshop on Definition Extraction*, RANLP-09. 18h September 2009, Borovets, Bulgaria, pp. 61-67.

Witten, I. H. and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: Elsevier.