Automatic Acquisition of Semantics-Extraction Patterns from Parallel Texts

Pavel Smrž

Faculty of Information Technology
Brno University of Technology
Božetechova 2, 61266 Brno
Czech Republic
smrz@fit.vutbr.cz

Abstract

This paper examines the use of parallel and comparable corpora for automatic acquisition of semantics-extraction patterns. It presents a new method of the pattern extraction which takes advantage of parallel texts to "port" text mining solutions from a source language to a target language. It is shown that the technique can help in situations when the extraction procedure is to be applied in a language (languages) with a limited set of available resources, e.g. domain-specific thesauri. The primary motivation of our work lies in a particular multilingual e-learning system. For testing purposes, other applications of the given approach were implemented. They include pattern extraction from general texts (tested on wordnet relations), acquisition of domain-specific patterns from large parallel corpus of legal EU documents, and mining of subjectivity expressions for multilingual opinion extraction system.

1. Introduction

Several text-mining, information-extraction or ontology-acquisition frameworks have been developed in the last decade (see, e.g., GATE (Bontcheva et al., 2002), OntoLT (Buitelaar et al., 2003), Mo'K Workbench (Bisson et al., 2000), OntoLearn (Gangemi et al., 2003), KnowItAll (Etzioni et al., 2004), Text2Onto (Cimiano & Voelker, 2005)). The employed methods are often based on an extension of the basic pattern-based extraction technique (Hearst, 1992) combined with token co-occurrence computation. Various modifications of the generic method are discussed in (Pantel et al., 2004).

The key aspect of the acquisition procedure is the definition of extraction patterns. The patterns can be specified either manually (with the knowledge of the particular language) or "learned" from data with the help of previously tagged semantic relations. Both approaches entail a repetition of the work for every new language in a multilingual environment.

This paper presents a new method of pattern extraction which takes advantage of parallel texts to "port" text mining solutions from a source language to a target language. It helps in situations when the extraction procedure is to be applied in a language (languages) with a limited set of available resources, e.g. domain-specific thesauri. For example, consider the current state of European legislation. There are legal dictionaries and thesauri for English, French and other "old-European" languages. However, these resources are also needed for new member (and candidate) states. Then, the process of the translation can be supplemented by our automatic methods.

The rest of the paper is organized as follows. The next section briefly discusses the context of our research – the area of multilingual technology-enhanced learning systems. Section 3 presents the method based on parallel or comparable texts in source and target languages. Section 4 introduces various applications that were used to evaluate (parts of) the implemented system. Section 5 indicates future directions of our research.

2. Multilingual pattern extraction for e-learning applications

The pattern extraction techniques as well as the multilingual knowledge mining system based on the acquired patterns presented in this paper are developed in the broader context of technology enhanced learning systems (Smrž & Nováček, 2006). We currently participate in a large project which aims at enabling a context-aware access to learning objects equipped with semantic descriptions. The knowledge objects are given in various languages and the system needs to implement common services for enriching and restructuring the learning material.

The patterns resulting from the described methods are directly applied for ontology acquisition and refinement of knowledge extracted from learning objects. Acquired semantic relations are used in various tasks:

1. Structuring of knowledge artefacts and their presentation to learners in the form of extracted ontologies. The user can also browse knowledge sources by means of the links that are automatically added to the stored learning objects.
2. Mining significant semantic relations helps to classify the content of knowledge objects in the repository. This is important especially for very narrow subfields with a limited number of documents that can be applied for standard text classifier training.
3. A definition of a search context is another area which benefits from automatically acquired semantic relations. Users of our system can restrict the search for (parts of) learning objects reflecting given semantic relations.
4. Last but not least, extracted knowledge based on the automatically acquired patterns enables personalized access to the knowledge artefacts in the repository. The learning management system can employ semantic relations to locate relevant learning material for specific needs of each particular user. Based on the user profiles, the system defines rules to identify "the best" information in the given situation.

type of the relation	subject	Object	relevance
participation_in	Iceland	Community Civil Protection Mechanism	0.85
when_implemented	CECIS	2004	0.82
is_a	Germany	member state	0.79
abbr_means	MIC	Monitoring and Information Centre	0.76
abbr_means	RAS	Rapid Alert System	0.62
is_a	virulence	priority criterion	0.45
is_a	smallpox	Agent	0.45

Table 1: An example of semantic relations extracted from the AC corpus

3. Mining parallel texts

The well known Zipf law holds not only for words, senses and other linguistic phenomena, but also for the availability of lexical resources across languages. There are very few languages with all the needed corpora, dictionaries, analysers, etc. and very many languages that lack even the basic set of them.

In this part, we discuss the situation when there are valuable resources for one language (English, in our case) that would be hard to create from scratch for other languages (Czech, Greek, …). At the same, there are large parallel corpora (for particular domains) that can help to "transfer" an important resource between languages. It is exactly the situation in the mentioned project – knowledge extraction patterns for specialized learning objects and other necessary resources are available for English and we can take advantage of English-Czech parallel corpora developed in previous projects.

Before particular steps of the algorithm will be presented, let us mention that the presented method assumes automatic acquisition of extraction patterns provided that a set of expected results is given. There are many machine learning techniques that enable this task. The quality of the obtained patterns depends considerably on the level of linguistic annotation of the corpus in question. If there are no stemmers, lemmatisers, morphological analysers, POS taggers, chunkers or parsers available for the target language, one cannot expect perfect results. However, even the simplest lexical patterns based just on token co-occurrences provide valuable results that could not be collected without them. All the target Czech texts were annotated just by the available morphological analyser AJKA (Sedláček & Smrž, 2001). We employed no disambiguation of the resulting morphological tags. Yet, the implemented method extracted correct patterns in most cases. This confirms our previous findings (Kilgarriff et al., 2004) that neither the free word order of Slavic languages, neither their rich inflection morphology rule out the use of lexico-syntactic patterns.

Figure 1 shows the overall schema of the implemented system. The process of pattern acquisition can be divided into the following steps:

1. Semantically-related terms are extracted from one (source) side of a parallel corpus. This extraction can be based either on pre-existing patterns or on the patterns automatically acquired from sample data. Note that we aim at rather deep semantic relations (not just hyper/hyponymy, synonymy etc., see Table 1). If the source is annotated appropriately, the extraction patterns will benefit from syntactic features represented by the tags.
2. The alignment of the parallel texts is used to extract translation equivalents in the target language. We experimented with the alignment on the level of sentences and short paragraphs. It is obvious that more elaborate alignment is used more accurate results can be expected.
3. Based on the target data specifying how the result of extraction should look like, the extraction patterns in the target language are derived and generalized over all the available data. This data does not need to be limited to the parallel corpus. Usually, it is advantageous to extract patterns from as much data as one can collect, even if it does not exactly match targeted domains. Again, the quality and the applicability of the acquired patterns are affected by the quality of corpus annotation.
4. The extraction patterns are applied to the particular data and a set of semantically related terms is automatically acquired.

The described procedure has been applied in three different environments (see Section 4). All the settings share the same preprocessing phase (for English). The system expects plain-text documents as the input. To reduce irrelevant data and increase the efficiency, the input is pre-processed. Pre-processing consists of splitting the text into phrases and eliminating irrelevant ones, tokenizing of the text, POS tagging and lemmatization, and chunking. The first two steps are based on regular expressions and are performed in one pass through the input file. The presence of potential semantically related terms is detected by matching "core words". The preprocessing tools take advantage of the NLTK toolkit (http://nltk.sourceforge.net) and GATE (http://gate.ac.uk). The designed modular architecture allows porting the system easily for different languages. After successful preprocessing, the extraction patterns are applied. Fast regular expression-based chunking is performed on the tagged phrases. Based on statistics computed from the pre-processed data, the most reliable patterns are extracted and a fuzzy assignment of their characteristics is performed (see (Nováček & Smrž, 2005) for details).

4. Applications of the extraction method

4.1. Wordnet- and thesaurus-based evaluation

To evaluate the multilingual system one needs to define source and target sets of semantic relations that should be automatically extracted based on the generated

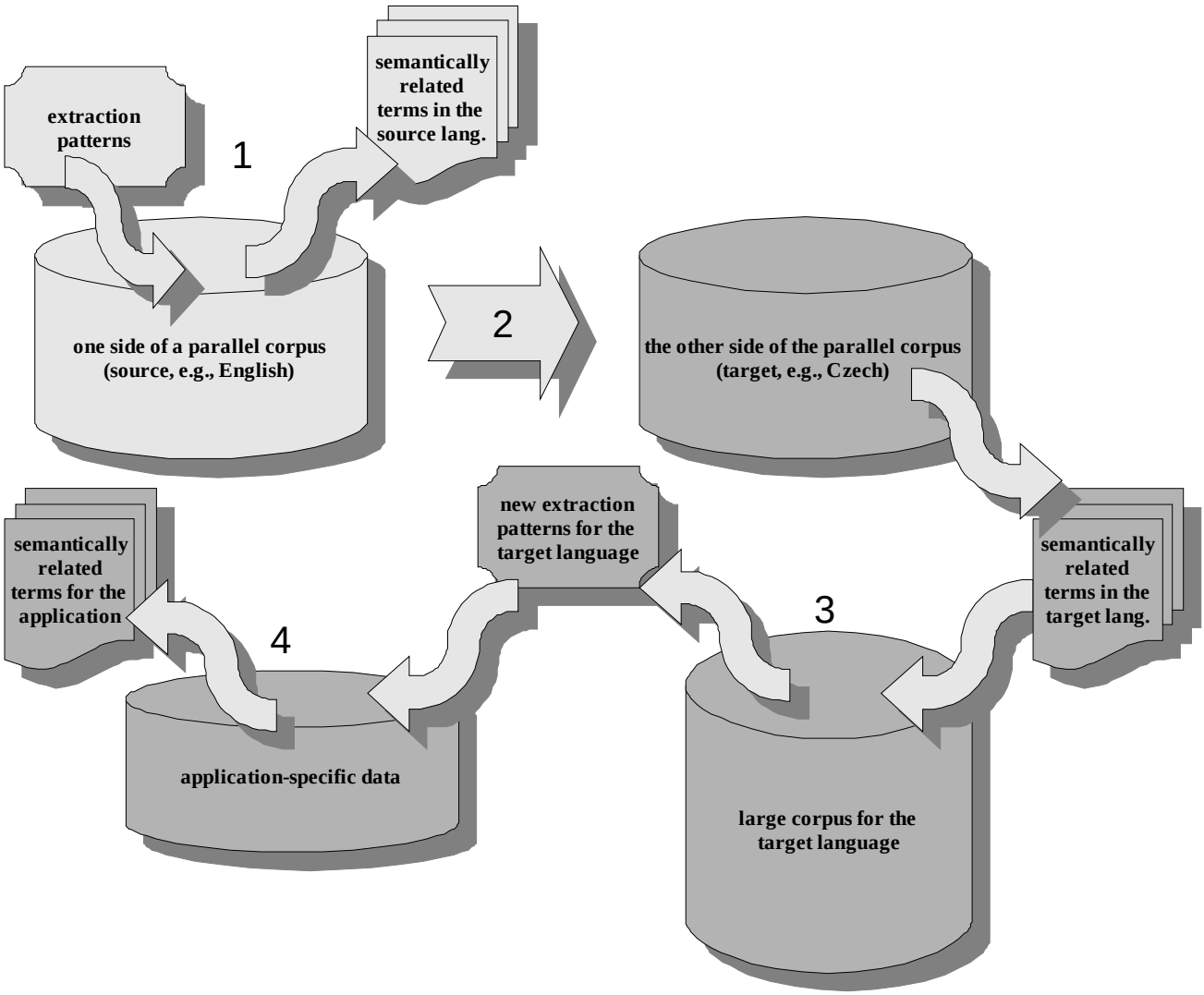

Figure 1: The basic schema of the patter extraction system

patterns. Moreover, it is necessary to run the experiments on large portions of parallel texts. We took advantage of the available resources from two previous European projects – EuroWordNet and BalkaNet – and compared the results on a subset of the relations from the standard Princeton wordnet that is covered in the national wordnets in other languages. This experiment proved that the automatically generated patterns enable extracting large percentage of the relations in the target language (78% for Czech). Obviously, the results depend on the nature of the target relations (from wordnets) and the type of (parallel) texts (general English-Czech corpus containing a lot of fiction).

We can also provide preliminary results of our system on a very large parallel corpus. The AC (Acquis Communautaire) corpus (http://wt.jrc.it/lt/acquis/) has been collected by the Joint Research Centre, Ispra, Italy. It contains thousands of legal EU documents covering a large variety of subject areas in 20 languages. The current experience with the application of the Eurovoc thesaurus (http://europa.eu.int/celex/eurovoc/) clearly shows that the basic method can bring valuable results. However, it is also obvious that the domain-specific corpus (European legislation) needs much more detailed analysis of the source texts and a deep semantic representation of the relations between concepts. This is also one of the directions of our future research.

4.2. Subjectivity clue mining

The described pattern acquisition algorithm has been also used for collecting patterns that can extract subjective expression from texts (subjectivity clues). The ultimate goal of our work in this area is to develop a multilingual system that analyses various information sources – news streams, blogs, forums – identifies and collects opinionated texts, and reports the diversity of the opinions on the same issue across countries and languages (see (Smrž, 2006) for details). Opinion mining comprises segmentation of documents, passages, sentences, or phrases to objective (factual) and subjective parts, and evaluation of the subjective attitude toward a given fact. Due to the long tradition of the research on the identification and extraction of private states (general term for opinions, emotions, sentiments, speculations, etc.) in English, there are large collections of the subjectivity

clues available for that language. We decided to "port" English patterns to Czech.

There is currently no parallel English-Czech corpus that would include enough data for the sentiment analysis. We used a special web crawler that searched for opinionated news articles in Czech that can be identified as translations of English originals. This data were further extended by automatically downloaded Czech news and their English equivalents provided by the Czech News Agency (http://www.ctk.cz/english/services).

The subjectivity clues for English were automatically extracted from the MPQA Opinion Corpus version 1.2 (Wiebe, 2005) by means of the method described in (Riloff, 2005). The collected parallel (comparable) texts were aligned by Moore's technique (Moore, 2002) to serve as the bridge between the clues in English and Czech. The transformation algorithm produced a set of extraction patters that were applied to testing data from Czech newspapers. We could not evaluate the results automatically as there is no gold standard for subjectivity analysis in Czech. However, the manual inspection of the extracted subjective sentences and their classification proved that the extracted patterns enable accurate analysis of opinionated text. A good deal of the errors was due to incorrect tagging. Thus, the accuracy should improve when a better POS tagger will be incorporated.

5. Conclusions and future directions

The presented method helps to convert lexico-syntactic acquisition patterns across languages. It is crucial especially for multilingual applications dealing with many languages that need to go beyond manually-tuned language-specific extraction. One of the areas where this approach can considerably accelerate the development of new tools is the field of technologically enhanced learning.

As mentioned above, the evaluation of our experiments suggests several direction of our future research. We will focus on a more elaborate semantic tagging of the Acquis Communautaire corpus aiming at a knowledge extraction procedure which is not limited to the current thesaurus-based relations.

Czech as a representative of Slavic languages with complex morphology and syntax (due to the free word order) pose a challenging problem for pattern-based information extraction. On the other hand, there are many language resources and tools available for our language. The future work will explore the dependence of the pattern quality on the level of annotation. We have to deal with the lack of language resources for many languages and it is important to know what results can be expected in such cases.

Subjectivity analysis and opinion mining in the multilingual context is another key component of our research. We currently combine standard text-based subjectivity clues with phonetic features that can help to analyse spoken data. The future evaluation procedure for the subjectivity analyser in Czech should therefore reflect the additional modality.

Finally, we will carry on the research in learning management systems. Partners in the mentioned project will define own evaluation procedures to assess our knowledge mining solution and the quality of the extraction patterns in their specific applications.

6. Acknowledgements

This work was partly supported by the Ministry of Education of the Czech Republic, Grant Project MSM 6383917201, and by the European Social Fund, project CZ04.1.03/3.2.15.1/0146.

7. References

Bisson, G., Nedellec, C., Canamero, L. (2000). Designing clustering methods for ontology building – The Mo'K workbench. In *Proc.of the ECAI Ontology Learning Workshop*, pp. 13–19.

Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H. (2004) Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*. 10(3/4), pp. 349-373.

Buitelaar, P., Olejnik, D., Sintek, M. (2003) OntoLT: A Protégé plug-in for ontology extraction from text. In *Proc. ISWC*, pp. 3-4.

Cimiano, P., Voelker, J. (2005). Text2Onto – A framework for ontology learning and data-driven change discovery. In *Proc. NLDB*, pp. 227-238.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., Yates, A. (2004). Web-scale information extraction in KnowItAll: Preliminary results. In *Proc. WWW 2004*, ACM Press, New York, NY, USA, pp. 100–110.

Gangemi, A., Navigli, R., Velardi, P. (2003). Corpus driven ontology learning: A method and its application to automated terminology translation. *IEEE Intelligent Systems*, pp. 22–31.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. COLING*, ACL, Morristown, NJ, USA, pp. 539–545.

Kilgarriff, A., Rychlý, P., Smrž P., Tugwell D. (2004). The Sketch Engine, In: *Proc. EURALEX*, pp. 105-116.

Moore, R. C. (2002) Fast and accurate sentence alignment of bilingual corpora. In *Proc. AMTA*, Tiburon, California), Springer-Verlag, pp. 135-244.

Nováček V., Smrž P. (2005): OLE – A new ontology learning platform, In: *RANLP*, Borovets, BG, Incoma, ISBN 954-91743-1-X, pp. 12-16.

Pantel, P., Ravichandran, D., Hovy, E. (2004). Towards terascale knowledge acquisition. In: *Proc. COLING*, pp. 771-777.

Riloff, E., Wiebe, J., Willian, P. (2005). Exploiting subjectivity classification to improve information extraction. *Proc. AAAI 2005*.

Sedláček, R., Smrž, P (2001) A new Czech morphological analyser ajka. In *Proc. TSD*, Springer-Verlag, ISBN 3-540-42557-8, pp. 100-107.

Smrž, P., Nováček, V. (2006) Ontology Acquisition for Automatic Building of Scientific Portals, In: *Proc. SOFSEM*, Springer, ISBN 3-540-31198-X, pp. 493-500.

Smrž, P. (2006) Using WordNet for Opinion Mining, In: *Proc. GWC*, MUNI, ISBN 80-210-3915-9, pp. 333-335.

Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Lang. Resources and Evaluation* 39 (2-3), pp. 165-210.