Topic ontology construction from English and Slovene language technologies corpora

Jasmina Smailović[1], Senja Pollak[1,2]

[1]Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
[2]Faculty of Arts, University of Ljubljana, Aškerčeva 2, Ljubljana, Slovenia
{jasmina.smailovic, senja.pollak}@ijs.si

Abstract

This paper presents the OntoGen topic ontology construction tool and the process of building topic ontologies from English and Slovene research papers in the domain of language technologies. We were interested in how cleaning the documents (e.g. removing the references section), manual concept moving and renaming, or using supervised active learning affect the ontologies.

Gradnja ontologij tematik iz angleškega in slovenskega korpusa jezikovnih tehnologij

V članku predstavljamo orodje OntoGen ter proces gradnje ontologij tematik iz angleških in slovenskih znanstvenih člankov s področja jezikovnih tehnologij. Zanimalo nas je, kako čiščenje člankov (npr. brisanje poglavja z viri), ročno preimenovanje in premeščanje konceptov ter uporaba metode aktivnega učenja vplivajo na ontologije tematik.

1. Introduction

Manual construction of taxonomies and ontologies represent a significant investment of human resources when used for modeling a new domain. Therefore, methods for (semi-)automatic extraction of domain knowledge from unstructured texts were developed. Automatic taxonomy construction was addressed in e.g. Navigli et al. (2011) and Kozareva and Hovy (2010).

While an *ontology* is a "formal, explicit specification of a shared conceptualization" (Gruber, 1993), represented as a set of domain concepts and the relationships between them, a *topic ontology* is a set of domain topics or concepts[1] – formed of related documents – represented by the most characteristic topic keywords and related by the *subconcept-of* relationship (Fortuna et al., 2005; 2007).

The task addressed in this paper is to semi-automatically construct a topic ontology from documents in the area of language technologies. This domain has already been modeled in previous research. The main domain publications were collected by Bird (2008). Joseph and Radev (2007) performed the citation analysis, the domain topic and trend analyses were done by Hall et al. (2008) and Paul and Girju (2009), using LDA (Blei et al. 2003).

In this work, our goal is to get an overview of the topics covered at the Slovene language technologies conference. To do so, we semi-automatically constructed two topic ontologies, one from papers written in Slovene, and the other from papers written in English. The constructed topic ontology is corpus-driven and represents only the concepts covered in the given corpus. For addressing this task, we used OntoGen[2] (Fortuna et al., 2005; 2007), a data-driven ontology editor, focusing on extracting and editing of topic ontologies. We investigated how the fact of removing information specific to scientific articles, like references or authors' names, affect the ontology. A very interesting part of this research was to compare the resulting topic ontologies with and without using supervised *active learning* (Cohn et al., 1994; Settles, 2009). We continue the research presented in (Smailović and Pollak, 2011), where the English articles were modeled into a topic ontology. In addition to the English topic ontology, in this paper we also model the Slovene part of the corpus as a separate topic ontology.

The paper is structured as follows. Section 2 describes the corpus, data preparation and the ontology editing tool. In Sections 3 and 4 the topic ontology construction process is presented. Section 5 provides the conclusions.

2. The corpus, data preparation and the OntoGen topic ontology construction tool

The articles for this case study were taken from the proceedings of the Slovene Language Technologies Conference (proceedings of seven conference editions are available online: http://www.sdjt.si/konference.html). As the papers (79 in English and 109 in Slovene) were available as PDF documents, we had to transform them into an appropriate textual format for the OntoGen tool, i.e. to the named-line document format. The first step was to transform the documents to a text-only document format. PDF to text conversion was performed, using the PDFBox[3] and Nitro PDF reader[4]. The text files were transformed to UTF-8 encoding. Next, we split the English and Slovene articles.

In this research, we present two settings, in the first one, the topic ontology is constructed without cleaning the documents and in the second one, semi-automatic data preprocessing is first performed. For the latter, using Perl scripts, we discarded parts of articles, such as authors' names, institutions, references, section numbers, tables, page numbers, etc. to get the "cleaned documents".

After presenting the documents in the named-line document format, OntoGen was used for building a topic

[1] In this paper words *concept* and *topic* are used as synonyms.
[2] http://ontogen.ijs.si/
[3] http://www.codeproject.com/KB/string/pdf2text.aspx
[4] http://www.nitropdf.com/

ontology. OntoGen is a semi-automatic and data-driven ontology editor. Semi-automatic means that the system is an interactive tool that aids the user during the topic ontology construction process. Data-driven means that most of the aid provided by the system is based on the underlying text data (document corpus) provided by the user. The system combines text mining techniques (*K*-means document clustering) with an efficient user interface to reduce both the time spent and complexity of manual ontology construction for the user.

3. Topic ontology on raw documents

We first examined how the topic ontology looks like if we do not perform any additional cleaning of the corpus. In this case, the text documents in the named-line format consist of an ID, followed by a title, names of the authors, main text of the article and references.

OntoGen uses *K*-means clustering, i.e. a method of cluster analysis which aims to partition *N* instances (documents, in our case) into *K* clusters in which each instance belongs to the cluster with the nearest mean. If we build a topic ontology automatically, by only suggesting to OntoGen the number *K* of concepts at each node of the concept hierarchy, the result for the English articles can be seen in Figure 1. For every concept, we tried different *K*-values and chose the one that splits the concept in the best way according to the user's understanding of the area. Neither the active learning functionality nor the renaming of the concepts was performed in this topic ontology construction process.

As one can see from the figure, names of the concepts/topics are not intuitive, and in some cases it is hard to understand what they represent. This happens since for concept naming OntoGen selects the first three most frequent words from the automatically constructed keywords list. For example, if the concept is described by the following keywords: *slovenian, translation, vowel, speakers, synthesis, speech, corpus, tagging etc.*, OntoGen will name this concept "*slovenian, translation, vowel*".

A better way of naming concepts is by involving the expert who can quickly find an appropriate concept name after observing all the topic keywords. Using this approach, the previous topic could be called *Speech technologies*. All the concepts in the English and Slovene topic ontologies were thus manually renamed based on the automatically extracted topic keywords.

Next, we observed that several topics/concepts were not present in the topic ontology. For the terms often occurring in the keyword lists of different concepts, but not being one of the three main topics keywords**,** we decided to use the supervised method for adding topics. It is based on the Support Vector Machine (SVM) active learning method of OntoGen. For the English corpus, we entered queries for *Speech recognition* and *Speech translation* concepts and answered some automatically proposed questions of a type: *Would you classify the document number 41 as an article on the topic of Speech recognition?* which enabled the system to label the instances. After the concept node was constructed, it was added to the ontology as a sub-concept of the selected concept, in our case, as a sub-concept of the *Speech technologies* concept. Similarly, we performed active learning also on the Slovene corpus. We entered queries for *Prevajanje govora* (*Speech translation*). In this way we tried to identify the most common and important words for the missing subconcept and put them in the query. Then, as for English articles, we answered several questions, and a new sub-concept was added in the Slovene topic ontology.

After manually renaming the concepts, using active learning for adding concepts, and manually moving some documents from one concept to another, we got an improved topic ontology. The resulting English topic ontology is shown in Figure 2. This ontology is more intuitive and understandable. One can see from the figure that language technologies consist of *Computational linguistics* and *Speech technologies* as its core concepts. This is also the general division of the field of language technology (e.g. in Wikipedia, language technology is defined as follows: *Language technology is often called human language technology (HLT) or natural language processing (NLP) and consists of computational linguistics (or CL) and speech technology as its core but includes also many application oriented aspects of them.*).

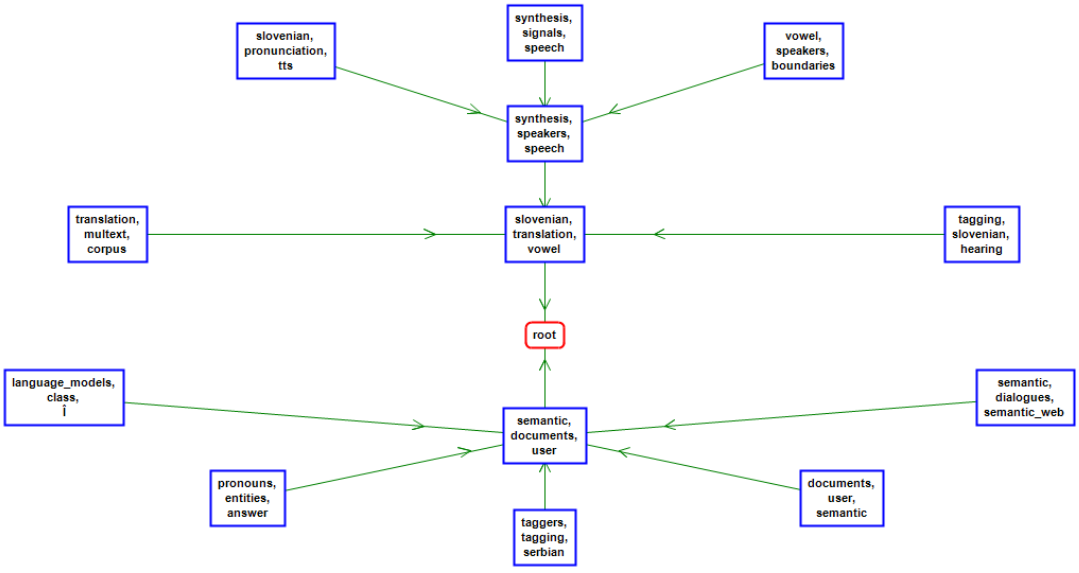

Fig. 1: English topic ontology without cleaning text document and without concepts renaming.

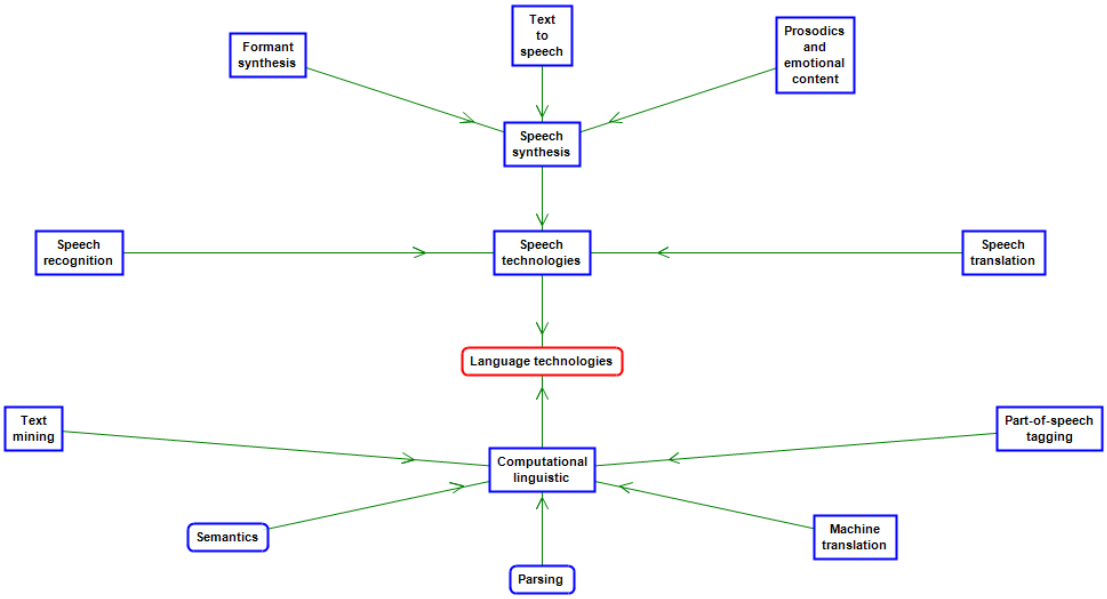

Fig. 2: English topic ontology after manually renaming the concepts, using active learning for adding concepts and manually moving some documents from one concept to another.

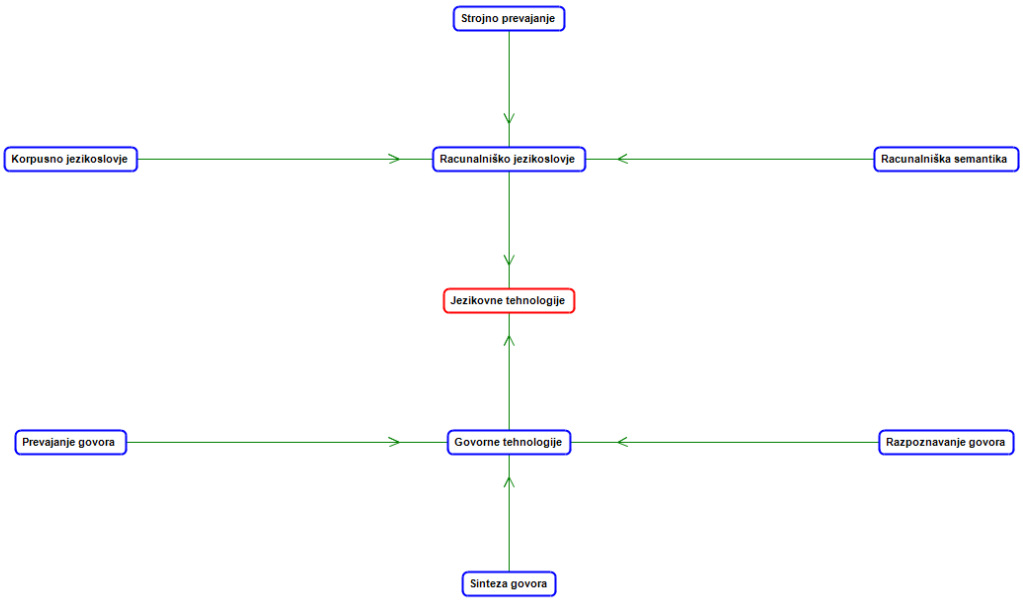

Fig. 3: Slovene topic ontology after manually renaming the concepts, using active learning for adding concepts and manually moving some documents from one concept to another.

Precise evaluation of the ontology coverage is a very hard task since we do not have a golden standard ontology for this specific corpus. We were therefore only able to approximately evaluate the coverage of the research area of language technologies separately for individual topics, as illustrated on the following subtopic. Concept of *Speech technologies* is in Wikipedia divided into 6 subfields (*Speech synthesis, Speech recognition, Speaker recognition, Speaker verification, Speech compression, Multimodal interaction*). The division in our ontology covers 2 out of these 6 concepts (and adds one more). However, as all the missing concepts occur very rarely in the corpus, the evaluation shows that OntoGen performs well, as the constructed ontology indeed adequately reflects the nature of the corpus.

Thus, OntoGen did the splitting very well for the root concept, we just had to change the sub-concepts' names. More manual work - supervised learning and manually moving some documents from one concept to another, needed to be done in further concepts splitting.

Slovene topic ontology, on the other hand, (after renaming the concepts, using active supervised learning and by manually moving some documents from one concept to another) is shown in Figure 3. Given that OntoGen does not have a stemmer for Slovene, we lemmatized the input documents in data preprocessing.

One can see from the figure that the Slovene topic ontology is simpler than the English one. For the Slovene topic ontology we had to do much more manual work (moving some documents from one concept to another). Interestingly, one third of the Slovene articles belong to

topic *Korpusno jezikoslovje* (*Corpus linguistics*). Note that the *Corpus linguistics* category in our ontology comprises also design of corpora and tools for corpus processing, such as PoS taggers and is not used in its strict sense, but in the sense of "Corpus construction and use".

Concept visualization for English articles can be seen in Figure 4. Two main concepts are marked with green and orange dashed lines. In the upper left corner of concept visualization one can notice some non-standard characters. These show OntoGen encoding problems probably due to the Slovene characters which are present in authors' names and references or some special characters from the tables.

Concept visualization for Slovene articles can be seen in Figure 5. The visualization is similar to the visualization for English articles, i.e. the articles are divided into two major topics (*Computational linguistics* and *Speech technologies*) and the *Computational linguistics* topic contains much more articles than the Speech technologies topic.

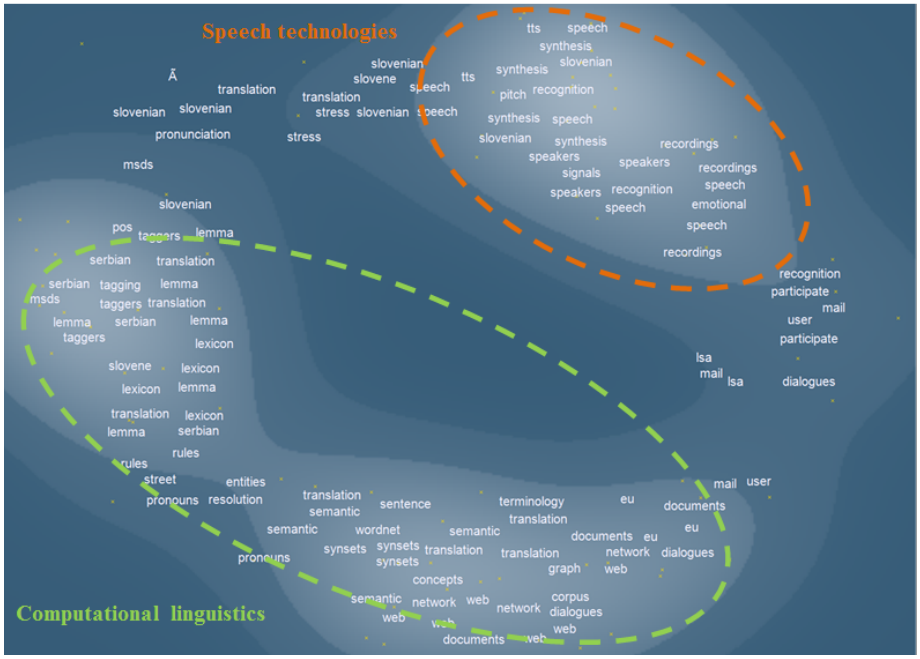

Fig. 4: Concept visualization of English text documents without cleaning (the automated splitting into two main topics, which we label *Computational linguistics* and *Speech technologies*, can be observed).

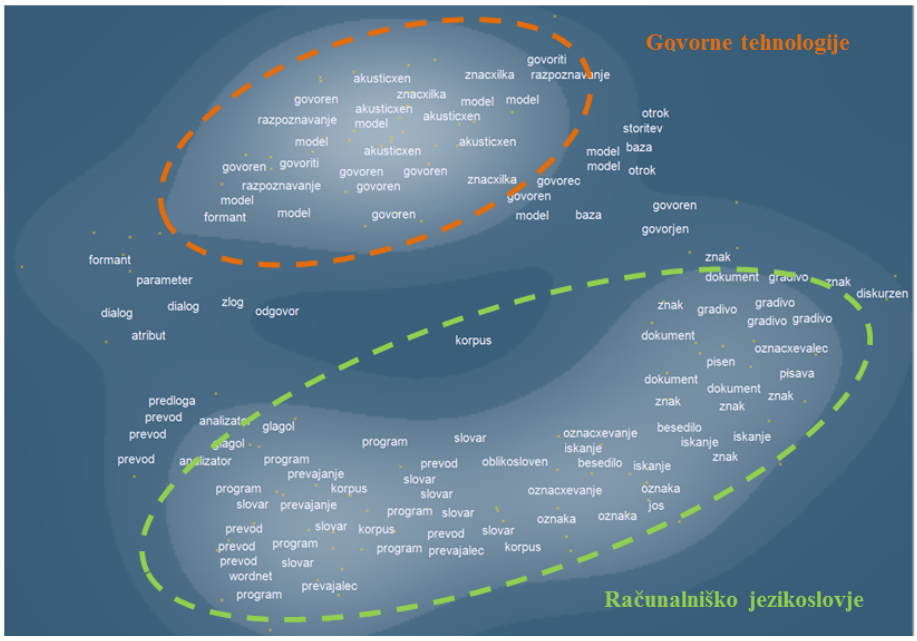

Fig. 5: Concept visualization from Slovene text documents without cleaning (the automated splitting into two main topics, that we label *Računalniško jezikoslovje* and *Govorne tehnologije* can be observed).

4. Topic ontologies constructed from "cleaned" corpora

We consider a text document to be "cleaned" once the names of the authors, references, page numbers, etc. were removed from the article.

In general, English topic ontology after cleaning the text documents has a similar structure as the topic ontology for text documents without cleaning (see Figure 6). Additional supervised learning and manually moving documents from one concept to another were also needed. One of the main differences is that the topic ontology for cleaned documents does not include the *Parsing* concept. We expected this type of differences, since the articles listed in the references may have an impact on the ontology, in our case for example, a new concept was created.

While building the Slovene topic ontology after cleaning the text documents (Figure 7), we noticed that splitting of the topic makes much more sense. Even the suggested topics` names were very similar to the actual names of the topics, even for leaf nodes of the hierarchy. For this ontology we did not perform any active learning or other manual work (except renaming), this is why the ontology is simpler than the topic ontology constructed from raw documents.

Concept visualization of English articles has visible differences, as shown in Figure 8. One can notice that once the text documents cleaned, certain groups of documents appear more distant. It is obvious that they were closer before because of the names of authors and names of papers and authors in the references. One can notice a non-standard character "Î" in the concept visualization. This character is also present after adding it to the stopword list, due to a mismatch with the encoding in OntoGen. After carefully reading the articles, we noticed that some of them contained Cyrillic characters which could not be properly encoded after PDF to text conversion. Concept visualization of Slovene articles is very similar to the one on uncleaned documents, but slight differences can be observed. In the *Računalniško jezikoslovje* (*Computational linguistics*) topic, one can see that documents which belong to concepts *Korpusno jezikoslovje* (*Corpus linguistics*) and *Strojno prevajanje* (*Machine translation*) are now distant (cf. Figure 9). Again, they were probably closer before because of the authors' names and the titles of papers and authors in the references.

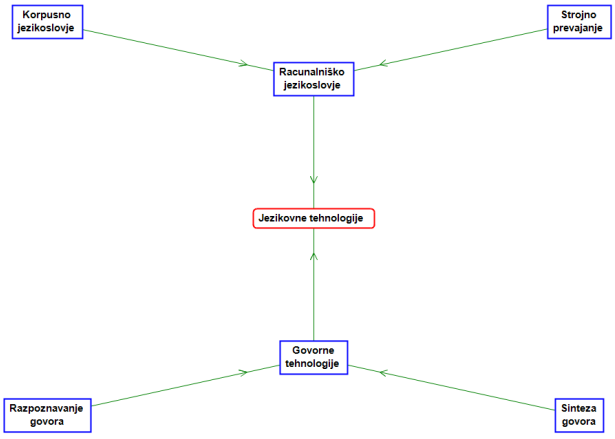

Fig. 7: Slovene topic ontology on cleaned text documents.

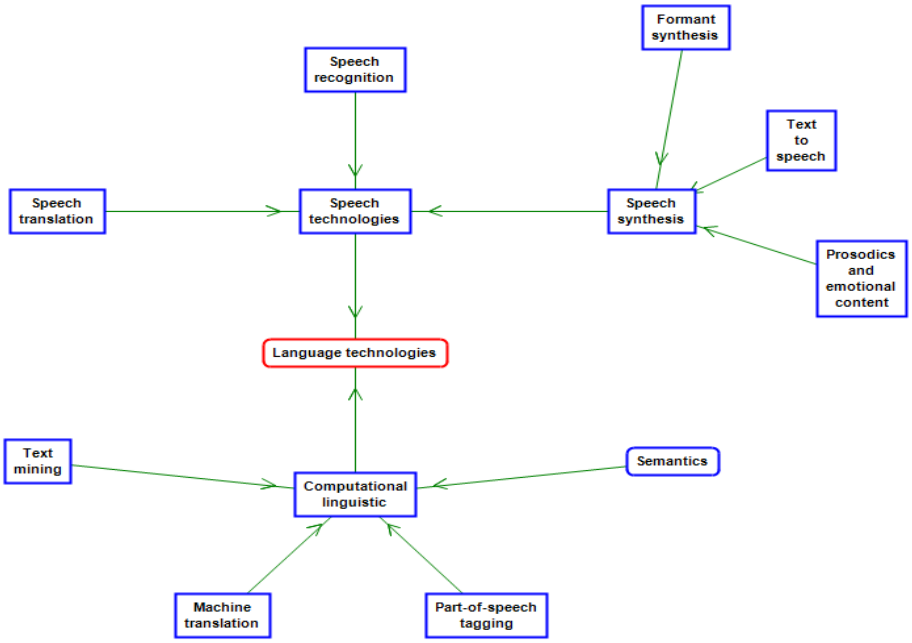

Fig. 6: English topic ontology on cleaned text documents after manually renaming the concepts, using active learning for adding concepts and manually moving some documents from one concept to another.

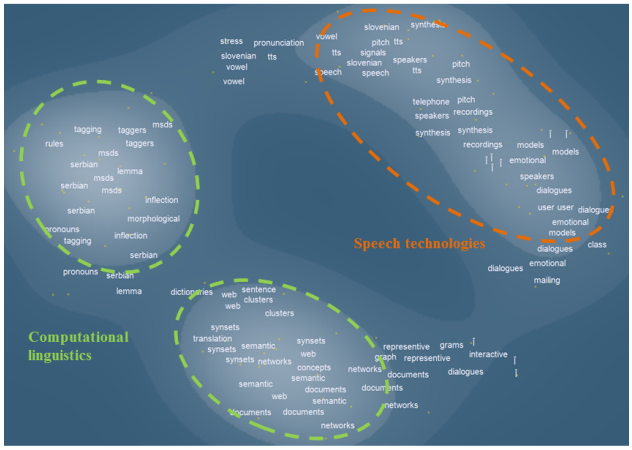

Fig. 8: Concept visualization of cleaned English articles marked with two main concepts.

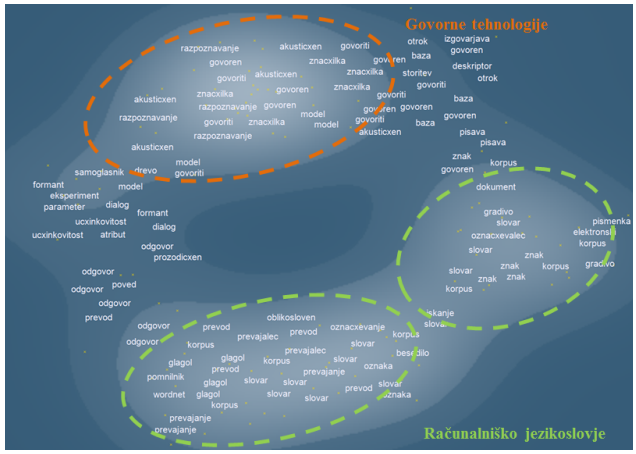

Fig. 9: Concept visualization of cleaned Slovene articles marked with two main concepts.

In other experiments, we also investigated how the topic ontology looks like if we use only the title and the abstract of the articles. However, we noticed that in the ontology building the splitting was done very well for the root concept, but further concepts splitting did not give satisfying results. Therefore, due to the space limitation, we do not present these topic ontology visualizations.

5. Conclusion

In this paper we have shown that OntoGen is a useful tool for constructing and editing topic ontologies, but as expected, to get perfect results additional manual work is necessary, using the OntoGen's semi-automatic editor.

We were interested in a specific domain of language technologies, and limited ourselves to the Slovene scientific conference in this field. Its small dataset (79 English and 109 Slovene articles) enabled us to evaluate the constructed ontologies, showing that adequate and interesting topic ontologies can be constructed. We extracted meaningful concepts from the language technologies domain and observed that the subtopics division (e.g. computational linguistics vs. speech technologies) make a lot of sense. This indicates that OntoGen could be used also as a tool for conference organization support, helping to assign articles to different tracks of the conference. Besides extracting the sub-concepts, we also identified hierarchical relations between them, which were not discussed in other studies. We can derive also some conclusions about data preprocessing. If one first cleans the documents by excluding authors' names and references from the document the cluster visualization reflects better the similarity between the articles contents.

6. Acknowledgements

We would like to thank T. Erjavec, N. Lavrač and B. Fortuna for their useful comments in the preparation of this paper.

7. References

Bird, S. (2008). Association for Computational Linguistics Anthology. http://www.aclweb.org/anthology-index/.

Blei, D., Ng, A. and Jordan, M. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

Cohn, D., Atlas, L. and Ladner, R. (1994). Improving generalization with active learning. Machine Learning, 15(2): 201–221.

Fortuna, B., Grobelnik, M. and Mladenić, D. (2005). Semi-automatic construction of topic ontology**.** Semantics, Web and Mining, Joint International Workshop, EWMF 2005 and KDO 2005, Porto, Portugal, Revised Selected Papers.

Fortuna, B., Grobelnik, M. and Mladenić, D. (2007). OntoGen: Semi-automatic ontology editor. In Smith, M. J. and Salvendy, G. (Eds.). In Proceedings of Human Interface, Part II, HCI International 2007, LNCS 4558, Springer, 309–318.

Gruber, T. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition. 5 (2):199–220.

Hall, D., Jurafsky, D. and Manning, C. (2008). Studying the history of ideas using topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), 363–371.

Joseph, M.T. and Radev, D.R. (2007). Citation analysis, centrality, and the ACL anthology. Technical Report CSE-TR-535-07, University of Michigan.

Kozareva, Z. and Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In Proceedings of EMNLP 2010, 1110–1118.

Navigli, R., Velardi, P. and Faralli, S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), 1872-1877.

Paul, M. and Girju, R. (2009). Topic modeling of research fields: An interdisciplinary perspective. In Proceedings of the International Conference RANLP-2009. Association for Computational Linguistics, 337–342.

Settles, B. (2009). Active learning literature survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Smailović, J. and Pollak, S. (2011). Semi-automated construction of a topic ontology from research papers in the domain of language technologies. In Proceedings of the Human language technologies as a challenge for computer science and linguistics (LTC'11), 121–125.