Corpus Based Extraction of Hypernyms
in Terminological Thesaurus for Land Surveying Domain

Vít Baisa, Vít Suchomel

Natural Language Processing Centre,
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xbaisa,xsuchom2}@fi.muni.cz

Abstract. We present two methods of automatic hypernym extraction implemented within a dictionary editing and writing system for terminological thesaurus for land surveying domain: a specialised corpus based method and a term similarity approach. Challenges of both techniques are briefly discussed.

Key words: thesaurus, terminological dictionary, corpus building, semantic relation, hypernym extraction, land surveying

1 Introduction

This work is an extension of dictionary editing and writing system for terminological thesaurus for land surveying domain [2]. The system consists of a terminology database and a text corpus. The terms are organized in a tree structure representing semantic relations hypernymy and hyponymy. The corpus is used for displaying the context of terms in domain-related documents (the concordance function), for finding synonyms[1] automatically (the thesaurus function) and extracting new terms (the term extraction function).

Having a terminology dictionary allowing terminologists to add their own documents to the underlying corpus and to extract new terms from the documents, a terminologist needs a function for placing the extracted terms into the existing tree structure of terms. In other words, the system has to provide a hypernym for every new term. Although there are methods for automatic hypernym finding (see below), the final decision must be made by a human expert. Therefore, it is useful to provide more hypernym candidates and let the terminologist decide which are true hypernyms or define the correct hypernym (or hypernyms) manually.

In this paper we present two methods for automatic hypernym extraction implemented within the terminological dictionary: a specialised corpus based method and a term similarity driven approach. Challenges of both techniques affecting the usability of hypernym candidates in the system are briefly discussed.

[1] More precisely, semantic similarity based on a shared collocational context.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015, pp. 1–6, 2015. © Tribun EU 2015

2 Related Work

The task of finding semantic relations between terms, especially hypernymy, is very similar to the task of finding definitions for terms. Once a definition (definiens) of a term is available, it is likely to contain a hypernym or synonyms of the given term (definiendum). Therefore this work has been inspired by [4] whose authors deal with finding English definitions in large corpora. A corpus is used as a source of definition texts. It is queried for patterns of words typical for definitions. We have exploited the same idea for extraction of hypernyms of Czech terms.

[3] reports low precision and recall for Polish, so the method may perform worse in morphologically richer languages than English.

Patterns of co-occurrences of words in semantic relations in Czech were studied in [5] and [6]. We have chosen the best patterns according to evaluations of those studies.

3 Extending the Specialised Corpus

The specialised corpus for land surveying and geo-information domain built in an earlier stage of the project [2] has been used. To improve the extraction of semantic relations, cadastre, land surveying, and geo-information related Czech laws and regulations were downloaded from the respective web sites and added into the corpus. Because of their purpose, these documents are likely to contain term definitions and terms in hypernym relations.

The corpus was augmented by the following texts obtained from the web portal of State Administration of Land Surveying and Cadastre (`http://www.cuzk.cz`) in July 2015:

- Overview of law regulations concerning land surveying and cadastre (24 documents).[2]
- Law regulations in land surveying and cadastre (13 documents).[3]
- Department regulations and actions (43 documents).[4]
- Data sets (5 documents).[5]
- INSPIRE (Infrastructure for spatial information in European Union) (14 documents).[6]

The extended corpus has been compiled and indexed for fast search in corpus manager Manatee/Bonito [1]. The current total size of the corpus is 12,691,252 positions (9,757,005 words including 3,864,481 nouns) in 27,389 documents.

[2] http://www.cuzk.cz/Predpisy/Prehled-pravnich-predpisu-souvisejicich-se-zememer.aspx

[3] http://www.cuzk.cz/Predpisy/Pravni-predpisy-v-oboru-zememerictvi-a-katastru.aspx

[4] http://www.cuzk.cz/Predpisy/Resortni-predpisy-a-opatreni/Pokyny-CUZK-1-15.aspx,
http://www.cuzk.cz/Predpisy/Resortni-predpisy-a-opatreni/Pokyny-CUZK-16-30.aspx,
http://www.cuzk.cz/Predpisy/Resortni-predpisy-a-opatreni/Pokyny-CUZK-31-42.aspx

[5] http://geoportal.cuzk.cz/%28S%28qn4tqoqg02oega1vqydc1o4g%29%29/Default.aspx?head_tab=sekce-02-gp&mode=TextMeta&text=dSady_uvod&menu=20&news=yes

[6] http://geoportal.cuzk.cz/%28S%28lyfis0b5dkkvrox2b5b1q2h2%29%29/Default.aspx?head_tab=sekce-04-gp&mode=TextMeta&text=inspire_uvod&menu=40&news=yes

4 Hypernym Extraction Methods

Two methods of automatic hypernym extraction were implemented within the terminological dictionary: a specialised corpus driven method and a term similarity based approach.

4.1 Corpus Based Extraction

The specialised domain corpus used for various functions in the system can also be exploited for hypernym extraction. The corpus is queried through the concordance API of Sketch Engine [1] for the following patterns in the CQL formalism[7]. The extracted hypernym candidates are sorted by log-Dice score:

$$\text{similarity} = \log_2\left(\frac{2 * \text{number of co-occurrences of both terms}}{\text{term 1 frequency} + \text{term 2 frequency}}\right)$$

Pattern 1: The hyponym + **is/are** *(je/jsou)* + the hypernym.[8]

```
2:[k="k1"&c="c1"] ([lc=","] [k="k1"])*([lc="a|i|nebo|či"] [k="k1"])?
[lemmalc="být"&tag="k5eAaImIp3.*"&lc!="ne.*"]
([k="k1"&c="c[1246]"] [k="k2"]{0,2})? 1:[k="k1"&c="c[1246]"] within <s/>
```

Examples of terms in hypernymic relation extracted by Pattern 1:

- loxodroma ⊂ křivka
- teodolit ⊂ geodetický přístroj
- územní řízení ⊂ správní řízení

Pattern 2: The hyponym + **and/or another/other/similar** *(a/nebo další/jiný/ostatní/podobný)* + the hypernym:

```
2:[k="k1"] ([lc=",|a|nebo|či"] [k="k1"])* [lc="a|i|nebo|či|zejména|ani"]
[lemmalc="také|též|některý|nějaký|než"]?
[lemmalc="další|jiný|ostatní|podobný"]
([k="k1"&c="c[1246]"] [k="k2"]{0,2})? 1:[k="k1"&c="c1"] within <s/>
```

Examples of terms in hypernymic relation extracted by Pattern 2:

- elektronický teodolit ⊂ měřický přístroj
- hospodářský pozemek ⊂ pozemková držba
- mapový znak ⊂ kartografický vyjadřovací prostředek

Pattern 3: The hyponym + **is/are kind/type/part/example/way of** *(je/jsou druhem/typem/částí/příkladem/způsobem)* + the hypernym:

[7] Corpus Querying Language. Developed at the Corpora and Lexicons group, IMS, University of Stuttgart. The CQL as used in Sketch Engine is an extension to the original language.

[8] The hypernym token is labelled by 1 and the hyponym token is labelled by 2 in all CQL queries here.

```
2:[k="k1"&c="c1"] ([lc=","] [k="k1"])* ([lc="a|i|nebo|či"] [k="k1"])?
[lemmalc="být"&tag="k5eAaImIp3.*"&lc!="ne.*"]
[k="k1"&(lemmalc="druh|typ|část|příklad|způsob")]
1:[k="k1"&c="c2"] within <s/>
```

Examples of terms in hypernymic relation extracted by Pattern 3:

- ionosféra ⊂ atmosféra
- Morfometrie ⊂ kartometrie
- pozemek ⊂ zemský povrch

4.2 Term Similarity Approach

The other implemented method of finding hypernyms of a term is searching the term database. The given term is compared to all existing terms in the system database. The most similar terms are expected to be good hypernym/hyponym candidates. We have adopted Jaccard distance of bigrams of characters with threshold of 0.5 as the similarity measure of two terms:

$$\text{similarity} = \frac{|\text{term 1 bigrams} \cap \text{term 2 bigrams}|}{|\text{term 1 bigrams} \cup \text{term 2 bigrams}|}$$

Examples of terms in hypernymic relation found in the database using the similarity measure:

- absolutní tíhový bod ⊂ tíhový bod
- fáze Měsíce ⊂ Měsíc
- způsob ochrany nemovitosti ⊂ nemovitost

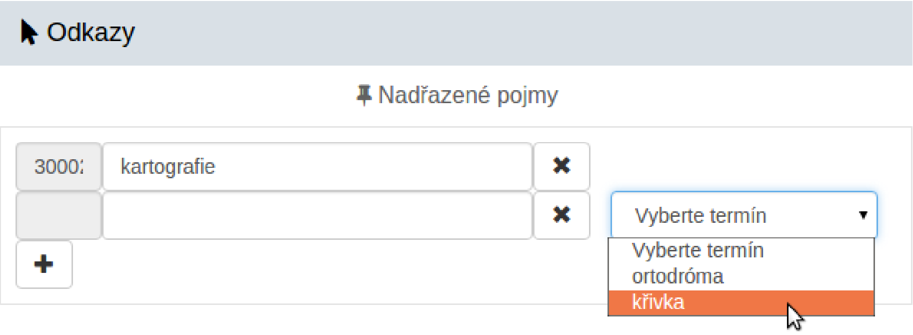
Both methods are combined in the system and the best candidates for hypernyms are available for the terminologists in the user interface in the form of a select box. An example of the (Czech) user interface is given in Figure 1.

5 Evaluation and Discussion

5.1 Corpus Based Extraction

The evaluation of top 25 and top 50 pair candidates extracted from the specialised corpus and sorted by log-Dice is shown in Table 1.

Although the accuracy of Pattern 1 and Pattern 2 queries is above 50 %, not all successfully extracted hypernym pairs are suitable for the particular term database. For example, term 'mapové dílo' ('map series') is a hyponym of 'dílo' ('series') however term 'kartografické dílo' ('cartographic product') – that is already in the term database – is a much more suitable hypernym.



3000: kartografie

Fig. 1. Automatic hypernym suggestions as implemented in the system. A part of editing form for the term 'loxodróma' is shown. Two terms visible in the select box are automatically suggested hypernyms.

Table 1. Percentage of hypernym pairs in all candidate pairs extracted from the corpus.

Pattern 1 – **'is'**		Pattern 2 – **'and other'**		Pattern 3 – **'is kind of'**	
candidates	hypernyms	candidates	hypernyms	candidates	hypernyms
top 25	52 %	top 25	52 %	top 25	0 %
top 50	56 %	top 50	60 %		

5.2 Term Similarity Approach

50 random terms from the database having at least one hypernym candidate were evaluated. The best three hypernym candidates were taken in account (the same as in the application). A hypernym was identified among the three most similar candidates in 56 % of cases.

Again, some successfully extracted hypernym pairs might not be suitable for the particular term database, because of e.g. a level in the hypernym tree is skipped or added or there is a better hypernym which was not identified by the automatic method.

That is why the final decision which candidate to select or whether to input the hypernym manually is left to a human expert.

6 Conclusion

We have successfully extended a dictionary editing and writing system for terminological thesaurus for land surveying domain by automatic hypernym extraction. The new function helps terminologists to keep the tree structure of the term database organised by suggesting candidates for hypernyms of terms.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and the research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

References

1. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: Ten Years On. Lexicography **1**(1) (2014) 7–36.
2. Horák, A., Rambousek, A., Suchomel, V., Kocincová, L.: Semiautomatic Building and Extension of Terminological Thesaurus for Land Surveying Domain. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2014. pp. 129–137.
3. Przepiorkowski, A., Degorski, L., Wojtowicz, B., Spousta, M., Kuboň, V., Simov, K., Osenova, P., Lemnitzer, L.: Towards the automatic extraction of definitions in Slavic. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies, ACL 2007. pp. 43–50.
4. Kovář, V., Močiariková, M., Baisa, V., Jakubíček, M.: Finding Definitions in Large Corpora with Sketch Engine. To appear. (Submitted to LREC 2016.)
5. Uhlíř, P.: Automatické vyhledávání nadřazených a podřazených pojmů v textu. Bachelor's thesis. Masaryk university. `http://is.muni.cz/th/255461/fi_b/`.
6. Haken, P.: Automatická extrakce sémanticky příbuzných slov. Bachelor's thesis. Masaryk university. `http://is.muni.cz/th/139525/fi_b/`.