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

Vít Suchomel, Vít Baisa

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

and
Lexical Computing Ltd.

5 December 2015

Partially supported by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

Generated by `paper2slides`.

Introduction

An extension of dictionary editing and writing system for terminological thesaurus:

- terminology database and a text corpus
- tree structure representing semantic relations hypernymy and hyponymy
- a specialised domain corpus used for concordance and term extraction

New: Two methods for automatic hypernym extraction:

- corpus based extraction
- term similarity driven approach

Related Work

- Definitions contain a hypernym or synonyms of the given term $\Rightarrow$ finding semantic relations between terms $\approx$ finding definitions for terms.
- Inspired by finding English definitions in large corpora.
- The method may perform worse in morphologically richer languages than English.
- Phrase patterns evaluated in MU students' theses dealing with co-occurrences of words in semantic relations in Czech were used.

Extending the Specialised Corpus

- Documents likely to contain term definitions and terms in hypernym relations needed.
- Cadastre, land surveying, and geo-information related Czech laws and regulations added.
- Source: Web portal of State Administration of Land Surveying and Cadastre (`cuzk.cz`).
- The corpus is encoded for fast search and loaded into Sketch Engine.

Table : Size of the augmented corpus

Documents	27,389
Positions	12,691,252
Words	9,757,005
Nouns	3,864,481

Corpus Based Extraction

- Three hypernym extraction patterns expressed as Corpus Querying Language queries.
- A term tokenised version of the corpus is queried through the concordance API of Sketch Engine.
- The extracted hypernym candidates are sorted by log-Dice score:

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

Hypernym CQL Pattern 1

The hyponym + **is/are** *(je/jsou)* + the hypernym.[1][2]

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

Examples of extracted pairs in hypernymic relation:

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

---

[1] Hypernym labelled by `1`, hyponym labelled by `2`.

[2] `lc`/`llc` = lowercased word/lemma, `k1` = noun, `k2` = adjective.

Hypernym CQL Pattern 2

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

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

Examples of extracted pairs:

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

Hypernym CQL Pattern 3

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

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

Examples of extracted pairs:

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

Evaluation – Corpus Based Extraction

Table : Top 25 and top 50 pairs of candidates sorted by log-Dice. Percentage of hypernym pairs in all candidate pairs extracted from the corpus.

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

Discussion: 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.

Term Similarity Approach

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.

Similarity measure: Jaccard distance of bigrams of characters with threshold of 0.5:

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

Examples of pairs found in the database:

- absolutní tíhový bod $\subset$ tíhový bod
- fáze Měsíce $\subset$ Měsíc
- ochrana nemovitosti $\subset$ nemovitost

Evaluation – Term Similarity Approach

Evaluation:

- The best three hypernym candidates of 50 random terms from the database having at least one hypernym candidate.
- A hypernym was identified among the three most similar candidates in 56 % of cases.

Discussion: 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.

User Interface

- Both methods are combined in the system.
- Up to 3 + 3 hypernym candidates given.
- The final decision which candidate to select or whether to input the hypernym manually is left to a human expert.

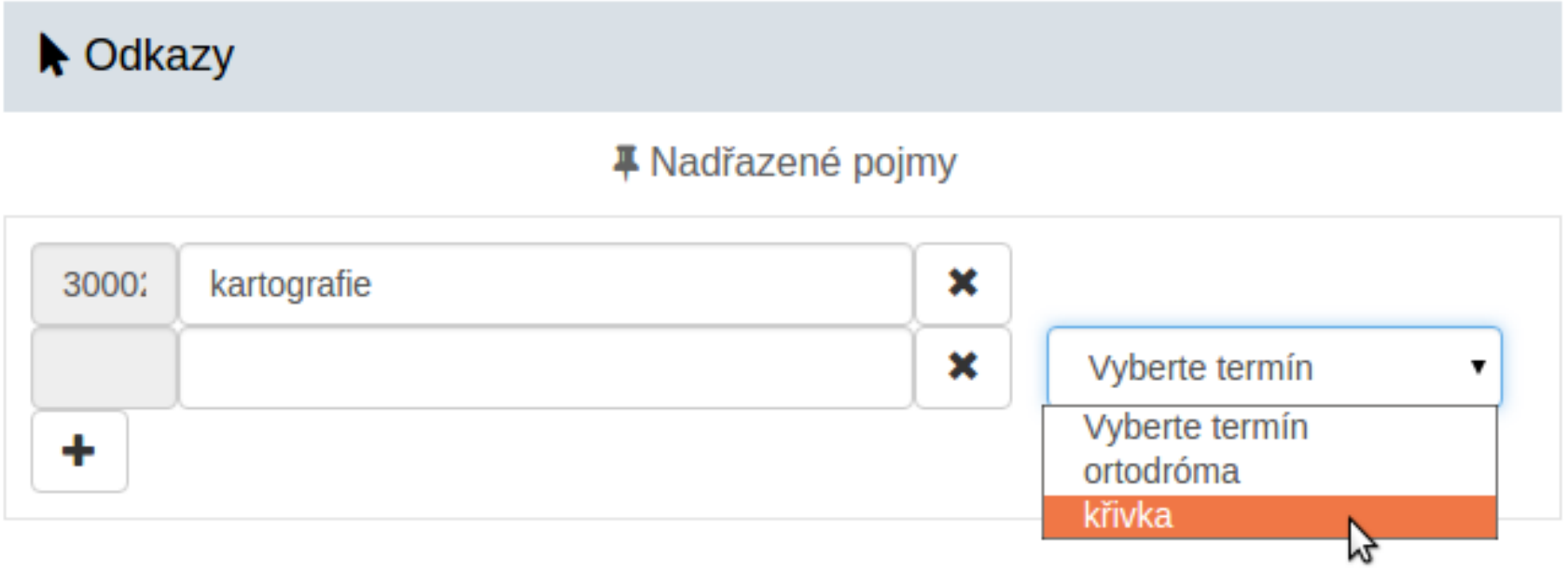

Figure : Automatic hypernym suggestions for the term ‘loxodróma’.

Conclusion

- Automatic hypernym extraction in Terminological Thesaurus
  - corpus based extraction
  - term similarity driven approach
- Up to 3 + 3 hypernym candidates given.
- Not perfect – the expert decides.