Terminology Extraction for Academic Slovene Using Sketch Engine

Darja Fišer[1,2] Vít Suchomel[3,4] Miloš Jakubíček[3,4]

[1]Department of Translation
Faculty of Arts
University of Ljubljana

[2]Department of Knowledge Technologies
Jožef Stefan Institute

[3]Natural Language Processing Centre
Faculty of Informatics
Masaryk University

[4]Lexical Computing

RASLAN, 2 Dec 2016

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

KAS Project

Project Goals

- National basic research project Slovene Scientific Texts: resources and description (2016-2018)
  - Partners: Jožef Stefan Institute, Ljubljana Uni, Maribor Uni
- Research questions & methodology:
  - What is Slovene scientific writing like?
  - How to support scientific writing in Slovene?
  - Combining HLT and linguistic investigations

Kas Corpus

- pdfs from Open Science Portal
- 50.000 texts, 4 million pages, 1 billion words

KAS Terminology

Term Extraction

- Terminology mining
  - monolingual
  - bilingual (structure-based)
- Term variants detection
- Enable terminology management to the scientific communities

Term Analysis

- Analysis of term usage by scientific field & time
- Analysis of term maturity in scientific fields
- Measure interdisciplinarity based on term usage

SkE TermExtract

- contrastive approach for finding term candidates
  - focus corpus consisting from texts in the target domain
  - reference corpus against which the focus corpus is compared
- SkE selects only grammatically valid phrases

Unithood

- rule based & language dependent
- tests grammatical validity of a phrase
- term grammar describes grammatically plausible terms with regex on MSD tags & lemmas

Termhood

- contrasts candidate phrases with reference corpus
- simplemath statistic compares candidates' normalized frequencies focusing on less/more frequent phrases

Term Grammar for Slovene v1.0

- based on Russian by M. Khokhlova & Czech by V. Suchomel
- extended with term patterns for Slovene by N. Logar

Default attributes

- 12 default attributes
  - 10 MSD-based
  - 2 agreement-based (gen, num, case) for improved accuracy of term identification

Term patterns

- used for identification & rendering of term candidates
- noun & verb phrases up to length 4
- noun, adjective, preposition, conjunction, adverb & verb

Term patterns

- 44 term patterns
  - 4-grams: 22 patterns
  - 3-grams: 15 patterns
  - bigrams: 6 patterns
  - unigrams: 1 pattern

```
*COLLOC "%(1.lemma_lc)_%(2.lc)_%(3.lc)_%(4.lc)-x"
1:n 2:adj_gen 3:adj_gen 4:n_gen & agree(2,4) & agree(3,4)
#"Nc.*" "A.*g.*" "A.*g.*" "Nc.*g.*"
#metoda magnetronskega ionskega naprševanja
```

Figure: Example of a term pattern

Evaluation

Corpus selection

- terminology typically extracted for a limited domain but our main goal was to evaluate the term grammar
- using a heterogeneous corpus more suitable to highlight different characteristics & issues across several domains
- focus corpus: KAS PhD
- reference corpus: slTenTen

KAS PhD

- 700 theses, 150,000 pages, 53 mio tokens, 2000-2015
- Social > Technical > Natural > Biomedical > Humanities

Results

Candidate selection

- 1,000 top-ranking term candidates
  - 4-grams: 28 (2.8%) term candidates
  - 3-grams: 177 (17.7%) term candidates
  - bigrams: 795 (79.5%) term candidates

Evaluation procedure

- 3 steps
  - pattern productivity: which patterns have a good yield
  - unithood & structural accuracy: identify bugs in term grammar
  - termhood: suggest refinements of term ranking & smoothing

Results for 4-grams

Pattern productivity

- 68%: noun phrases with preposition

Structural accuracy & unithood

- 75%: candidates with prepositions or conjunctions
- 40%: adjective & noun combinations (term rendering)
- unithood problems: truncated candidates

Termhood

- 75%: candidates with conjunctions
- 60%: candidates with prepositions
- 58%: adjective & noun combinations
- false positives: general-language (`pogovor o likovni nalogi`) & thesis-specific (`cestni otrok v makejevki`)

Results for 3-grams

Pattern productivity

- 43%: adjective & noun combinations
- 42%: candidates with prepositions

Structural accuracy & unithood

- 83%: candidates with prepositions
- 70%: adjective & noun combinations
- unithood problem: candidates subsumed in longer phrases (`sistem za podporo *odločanju*`)

Termhood

- 73%: adjective & noun combinations
- 63%: candidats with prepositions
- tech & nat sci: unithood, soc sci & hum: termhood

Conclusion

Summary

- substantially fewer 4-grams were extracted, but their pattern range was greater than in 3-grams
- lower unithood & structural accuracy in 4-grams
- termhood results similar in both
- term grammar refinement needed

Future work

- applicable to all corpora with compatible MSD tagging
- freely available: `http://nl.ijs.si/kas/english/`
- comparison of domain-specific subcorpora
- comparison of MA and BA theses
- comparison with CollTerm
- development of Slavic Term Grammar Core

Terminology Extraction for Academic Slovene Using Sketch Engine

Darja Fišer[1,2] Vít Suchomel[3,4] Miloš Jakubíček[3,4]

[1]Department of Translation
Faculty of Arts
University of Ljubljana

[2]Department of Knowledge Technologies
Jožef Stefan Institute

[3]Natural Language Processing Centre
Faculty of Informatics
Masaryk University

[4]Lexical Computing

RASLAN, 2 Dec 2016