First steps toward developing a system for terminology extraction

Petra Bago, Damir Boras, Nikola Ljubešić

Department of Information Sciences
Faculty of Humanities and Social Sciences
University of Zagreb

November 5, 2009

Introduction

- terminology extraction = (semi)automatic extraction of technical terms of a specific domain
- mostly systems that offer term candidates - lexicographer / lexicologist chooses from the list
- products - specialized dictionaries for human / computational use
- approaches:

1. statistical - differential analysis - finding differences between general and specialized corpora
2. linguistic - mostly on morphological and syntactic level
3. hybrid - combines the above mentioned methods

Building the data sample

- synopsis corpus
- documents downloaded from official web pages of the Faculty of Humanities and Social Sciences
- 420 synopses of doctoral theses
- time period: 2004-2009
- exclusively digital texts in .doc format
- manually structured into a relational database
  - title
  - introduction
  - theoretical background
  - narrower field of work
  - aims and problems of research
  - methodology
  - expected scientific and/or practical contribution
  - structure of thesis

Processing and analysis

- verticalization i.e. tokenization of synopsis corpus
- semi-automatic lemmatization - different morphological resources used on non-homographs, the rest lemmatized by hand
- two new columns: lemma of a particular token and its part of speech
- 420 documents
- fields: humanities 72.62%, social sciences 27.38%
- 338,706 tokens, 45,788 types
- average number of tokens per document: 806.44
- average number of types per document: 51.32
- type-token ratio: 0.135
- finance texts: 0.05, balanced corpora $\sim$ 0.1
- conclusion - diverse vocabulary
- language: Croatian language 95.08%, other languages 4,92%

Initial experiment

- goal of this research: get a feel for the data and the terminology extraction problem in general
- aspiration: develop a system for terminology extraction
- approach: data-driven - statistical - differential analysis of reference and domain-specific corpora
- experiment:
  - different reference corpora
    - large general corpus
    - small general corpus
    - specific corpus
  - linguistic pre-processing
    - using tokens
    - using lemmata
    - using POS filters

Building the gold standard

- a small sample which was manually annotated - tokens that are terms or part of multiword terms - one article - 671 tokens, 70 tokens tagged
- tagged by only one person - no inter-annotator agreement can be computed - no ceiling
- no possibility of having a development and an additional test corpus
- all plans for the future to make the methodology more accurate
- corpus verticalized
- additional columns containing:
  - a term or part of a multiword term (value 1), or not (value 0)
  - lemma
  - part-of-speech tag

Syntactic patterns

- most frequent simple patterns, highly complex patterns also occur
- example: "... postmodernom ili postindustrijskom, a kod nas i postsocijalističkom društvu..." (eng. "postmodern or postindustrial, with us also postsocialist society")

N	11
AN	8
A	4
NA(g)N(g)	3
ANCN	3
ACAN	2
AxAxxxxAN	1
NN(g)CN(g)	1
NN(g)	1

- more terms share a common head - decision - locating tokens

The log-likelihood ratio test

- statistical significance test introduced in (Dunning, 1993)
- corpus differential analysis - most popular
- compares two hypotheses
  - H_0 - tokens found in reference and specialized corpus are from the same distribution
  - H_1 - they are from two different distributions
- likelihood of a token computed through binomial distribution

$$L(p, k, n) = p^k(1-p)^{n-k}$$

- test statistic $-2\log\lambda$

$$2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)]$$

Reference corpora

- 3 different reference corpora

1. "Vjesnik1"
   - large newspaper corpus, on-line version of the daily newspaper Vjesnik
   - size: 746,683 tokens, lemma and part of speech
   - differently tagged than Synopsis coprus - TNT tagger (Agić, 2006)
2. "Vjesnik2"
   - subset of the large newspaper corpus
   - size: 70,000 tokens, lemma and part of speech
3. "Synopsis"
   - corpus described in this paper
   - size: 338,035 tokens, lemma and part of speech
   - all documents but the one used as the gold standard

Evaluation

- evaluation measures - precision, recall, $F_{0.5}$, F_1, F_2
- $F_{0.5}$ precision as twice as important as recall, F_2 the opposite
- F_2 considered optimal - all terms should be in the term candidate list - reviewed by an expert
- results for all three reference corpora will be shown, test statistic $-2 \log \lambda$ threshold optimized by maximizing the F_2 value
- baseline - random results - the result obtained by guessing

1. tokens without POS filtering - 70/671 - 10.43%
2. lemmata without POS filtering - 47/308 - 15.26%
3. tokens with POS filtering - 70/387 - 18.09%

First experiment

- tokens as features (baseline 10.43%)

RC	precision	recall	$F_{0.5}$	F_1	F_2	$-2 \log \lambda$
Vjesnik1	0.183	0.757	0.215	0.294	0.465	7
Vjesnik2	**0.194**	**0.757**	0.228	0.309	**0.479**	7
Synopsis	0.180	0.743	0.212	0.290	0.457	4

- POS distribution in reference corpus and result

part of speech	reference corpus	result	difference
noun	0.384	0.56	+45.8%
adjective	0.274	0.32	+16.8%
verb	0.101	0.10	-1.0%
other	0.242	0.02	-91.7%

Second experiment

- lemmata as features (baseline 15.26%)

RC	precision	recall	$F_{0.5}$	F_1	F_2	$-2 \log \lambda$
Vjesnik1	0.118	0.514	0.139	0.191	0.307	1
Vjesnik2	**0.125**	**0.600**	0.148	0.206	**0.340**	1
Synopsis	0.152	0.486	0.176	0.231	0.337	2

- POS distribution in gold standard and result of first experiment

part of speech	gold standard	percentage	result	percentage
noun	39	55.7%	144	53.9%
adjective	25	35.7%	84	31.5%
verb	0	0.0%	27	10.1%
other	6	8.6%	12	4.5%

Third experiment

- tokens as features, POS filter introduced (baseline 18.09%)

RC	precision	recall	$F_{0.5}$	F_1	F_2	$-2 \log \lambda$
Vjesnik1	0.220	0.813	0.258	0.347	0.528	7
Vjesnik2	**0.205**	**0.891**	0.242	0.333	**0.534**	5
Synopsis	0.211	0.859	0.248	0.338	0.532	3

- shows best results
- second experiment worst - through morphological unification many different features considered same, additional noise by error
- POS filter improves both recall and precision (91.4% of terms nouns and adjectives)

Further research

- a larger annotated sample (now only 671 tokens)
- different document sizes - important for differential analysis
- different text complexity
- samples annotated by more annotators - inter-annotator agreement
- methodology of using distinct development and testing samples
- experiments concerning the size and content of reference corpora
- include the minimum frequency criterion for document features
- experiment with more methods for differential analysis

Conclusion

- data sample of 420 documents and 338,706 tokens, high type-token ratio - complex vocabulary, syntactical complexity
- small gold standard built - has to be increased
- a smaller newspaper reference corpus yields better results than the big one - further research necessary - pure chance?
- using lemmata as document features - results consistently worse, loss of information greater than gain by morphological normalization (combining both features?)
- POS filter - improves F_2 significantly
  - without using the filter nouns and adjectives are chosen more often than by chance
  - their variation between corpora greater than of verbs and other parts of speech
  - nouns variate more than adjectives
- general conclusion: investigated methods achieve significantly better results than the random baseline