Domain independent Automatic Term Extraction Framework

Gintare Grigonyte
Saarland University
Saarbrucken, Germany
gintare@iai.uni-sb.de

Johann Haller
Saarland University
Saarbrucken, Germany
hans@iai.uni-sb.de

1 Introduction

Terminology extraction plays an important role in building lexical resources and currently is productively applied in IE, IR, ontologies and Knowlwdge Bases building fields. From NLP perspective, there are several approaches for terminology extraction: linguistic, statistic and hybrid. Terminology extraction systems based on linguistic approaches have a higher than 70% coverage in term extraction (see Bennet (1999), Borigault (2001)). Statistical term extraction approaches, with given big annotated training corpus, can perform almost the same good, but these methods do not always guarantee integrity and wholeness of the term (Frantzi (1999), Jisong Chen (2006)). The practice however shows that linguistic, i.e. rule based, approaches outperforms statistical ones in precision, and that combining linguistic and statistic approaches in various stages of term extraction process can benefit terminology extraction (Schiller (1996), Bourigat (2001)).

In the process of term extraction (Nagakawa (2001)) a "recognizing of all NPs" step (or so called extraction of term candidates) is considered a default. Since domain language is a more specific subset of a general language, general, rule based language processing tools can by applied for any domain terminology extraction. The important question is how to distinct domain specific terms from general NPs. Nagakawa (2001) notes, that in order to extract domain specific terms from term candidates, a ranking of term candidates according to their termhood is necessary. The term informativity (termhood) can be captured by statistical methods (IDF, MI, log likelihood, entropy, etc.).

Our approach to term extraction and building structured lexical resourses is based on linguistic patter matching for automatic term candidates extraction and IDF measurement for term quality assurance. It is in a way similar to approaches presented by Borigault (1992), Daille (1994), and Paulo (2002). However, the difference lays in working languages: for English and French there is no need to analyze compound words. This is though necessary for German language.

2 The Methodology

We present domain independent hybrid term extraction framework for English and German languages, which has following conceptual/architectural layers:

1) Rule based **morphological analysis**. For each word in the text, system delivers information such as lemma, PoS, derivation, semantic class etc. For instance, word `Anschlagsziel` (En. attack target) is analyzed like:

```
{string=Anschlagsziel,
lu=anschlagsziel,c=noun,ehead=
{nb=sg,case=acc;dat;nom,g=n},
gra=cap,cs=n#n,
ts=anschlags#ziel,
t=anschlag#ziel,
ds=an_$schlagen~IRREG#ziel,
ls=an_$schlagen#ziel,
ss=act#loc,s=loc}[1]
```

2) Rule based **disambiguation and syntactic analysis**. We use KURD[2] - a formalism that interprets rules based on finite-state technology. An example rule for identifying NP:

```
noun_phrase =
^Ae{c=w,sc=art,ehead=_AGR},
        *Ae{c=adj,ehead=_AGR},
         Ae{c=noun,ehead=_AGR}
      :  Au{ehead=_AGR}g{c=np}.
```

[1] A comprehensive explanation of tags can be found at: http://www.iai.uni-sb.de/docs/mmpro.pdf
[2] http://iai.iai.uni-sb.de/~carl/fred.pdf

3) **Variant and non-basic term form detection** is possible due to a detailed morphological analysis:

```
Atom dis_$place~ment
Atom~ic dis_$place~ment
```

4) **Stop words and filtering**. We use a general stop word list that contains words like: `less, never, next,` etc. However, after extracting candidate terms, stop words list is filled with new stop words, that serve as a filter.

5) **Phrase marking**. Morphological and statistical analysis is followed by the tagging of acronyms, proper names, possible single word terms and noun phrases:

```
Based on the <style code= "acro-
nym">VFF</style>
approach, an <style code="simpl">
approach </style>
to find the <style code="np"> opti-
mal number </style>
[...]
```

6) **Candidate term extraction**. Term extraction can serve for IR, IE, ontology learning and other knowledge acquisition from text tasks. Combining rich morphological and syntactical analysis with pattern matching NPs recognition, helps to extract a wide span of entities:

```
Possible Terms: software fault; re-
dundant system;
Toponyms: England;
Acronym: SCHEME;
Names of Persons and Organizations:
Jack Goldberg; N. Levitt; John H.
Wensley Computer Science Group;
```

7) **Statistical term informativity measure**. IDF measure is used. For each term t where $|D|$ is the number of all documents in the collection and d a single document from the collection:

$$idf(t) = \log\left(\frac{|D|}{\{d : t \in d\}}\right)$$

8) **Hierarchical representation building**. Extracted terms are represented via hypernym-hyponym relationship. To create a hierarchy from general to more special terms we used a simple method: non-compound terms are top level hierarchy nodes; for a term t_x with n compound parts, we look up whether there is a term t_y consisting of the n-1 rightmost term parts; if so, the term t_x becomes a subterm of t_y.

3 The experiment

The text resources used in the experiment cover approximately 2500 abstracts (in English) of papers in Dependability and Security domain. The corpus contains 181,548 tokens. Processing from step 1 to step 6, we gained 6818 terms. After the informativity values have been obtained, and we have defined certain threshold, the term list was pruned down to 5,710. All the steps were fully automated. Evaluation of the system (a sample of 10% of the abstracts) showed 82% of recall (in terms of IE system measurements, which would be called 18% of silence in the term extraction field) and 67% of precision (noise=33%). After applying IDF filter precision increased up to 79% (noise decreased to 21%).

References

Bennet N. 1999. *Extracting Noun Phrases for all of MEDLIN.* In Proc. Am. Medical Informatics Assoc. Symp., AMIA.

Borigault D., Jacquemin C., and M.-C. L'Homme, editors. 2001. *Recent Advances in Computational Terminology.* John Benjamins Publishing Company.

Bourigault, D. 1992. *Surface grammatical analysis for the extraction of terminological noun phrases.* In Proceedings of COLING-92, 977-981.

Daille, B., Gaussier, E., and Langé, J.M. 1994. *Towards automatic extraction of monolingual and bilingual terminology.* In Proceedings of COLING-94., 515-521.

Frantzi K. and Ananiadou S.. 1999. *The c-value/nc-value domain independent method for multiword term extraction.* Journal of Natural Language Processing, 6(3):145–179

Jisong Chen. 2006. *A Multi-word Term Extraction System.* PRICAI 2006: Trends in Artificial Intelligence, Springer, 1160-1165

Nakagawa, H. 2001. *Experimental Evaluation of Ranking and Selection Methods in Term Extraction.* In: Recent Advances in Computational Terminology. John Benjamins Publishing Company, 303-325.

Paulo, J. L. et al. 2002. *Using Morphological, Syntactical, and Statistical Information for Automatic Term Acquisition.* In E. Ranchhod and N. Mamede (eds.), Advances in Natural Language Processing, PorTAL. Springer-Verlag, LNAI 2389: 219-227

Schiller, A. 1996. *Multilingual Finite-state noun phrase extraction.* In: Proc. ECAI-96 Workshop on Extended Finite State Models of Language