A Hybrid Technique for Automatic Term Extraction

Byron Georgantopoulos (*)(@) and Stelios Piperidis (*)(&)

(*) Institute for Language and Speech Processing - ILSP
Artemidos & Epidavrou, Athens, GREECE
Tel: +30 1 6875300, fax: +30 1 6854270

(@) University of Athens

(&) National Technical University of Athens

Email: {byron, spip}@ilsp.gr

Abstract

In this paper we present a method aiming at (semi-)automating the process of eliciting domain specific terminological resources. The method aims at linguistically processing machine-readable text corpora and extracting lists of candidate multi-word terms of the domain, that would then be validated by domain experts. The method proceeds in three pipelined stages: a) morphosyntactic annotation of the domain corpus, b) corpus parsing based on a pattern grammar endowed with regular expressions and feature-structure unification, c). statistical evaluation of the candidate terms with an aim to skim valid domain terms and lessen the overgeneration effect caused by the pattern grammar. This hybrid methodology was tested on a Greek software manual corpus, featuring a 63% recall. Out of 10 different statistical filters applied on two-word terms, the best performing one further confirmed 30% of the index two-word terms and also reduced the size of the proposed list to 1/15.

1. Introduction

In this paper we present a hybrid method for automatically extracting terminological resources from text corpora. Automatic term extraction is of great interest nowadays, where huge volumes of texts are produced and published electronically, resulting in new requirements for their management and processing (automatic classification, information retrieval, information extraction, etc.). The application of language technology systems in order to meet these requirements presupposes the customisation of the system to the domain of the processed text. A basic step towards this procedure is the improvement and enrichment of language resources by incorporating the appropriate domain terminology. The application of methods for automatic extraction of terms provides a valid, fast and low-cost solution to the customisation procedure.

There are many natural language processing applications that automatic term extraction supports:

- text indexing - extracted terms construct a back-of-the-book index
- information retrieval - user queries are answered after comparing the terms of the query with the terms of the texts in the collection
- text classification - texts with equal or similar term sets are classified in the same domain
- text condensation / draft summarisation - sentences containing dominant terms can be considered as highly representative of the content of the original text. The set of terms itself can serve as a kind of mini-abstract.
- bilingual text alignment - terms of one language usually are translated uniformly in another language, within the same domain.

2. Methodological approaches

A term is a linguistic realisation of a domain specific concept and usually is lexicalised in the form of a noun phrase. In bibliography, one can find two basic methods for extracting terms:

1. Using a term grammar (usually a context free grammar) which is applied to an appropriately annotated text and extracts all the phrases it recognises [1]
2. Using statistical tools similar to the ones developed in the field of information retrieval and text indexing.
These tools include frequency counting, formulas from information theory, formulas which take into account the context of words, etc. [2],[9]

There are important differences between these two lines of action. A term grammar describes the syntactic structure that a valid term must satisfy, but it is possible that phrases recognised by the grammar are not valid terms. The weakness of a grammar is attributed to the fact that its rules, though a subset of NP rules, are general enough to generate a large number of potential terms1. Furthermore, a grammar cannot locate single word terms since such a term does not have any syntactic structure except part-of-speech information2. In general, a term grammar can only produce a set of potential terms that remain to be validated by an expert or a module of different nature.

The statistical approach is based on the assumption that words and phrases indicative of the domain of a document tend to appear frequently (the same applies for phrases consisting of words that appear frequently together). Frequency can have two different interpretations: (1) a phrase is more frequent in the current text than in a representative collection of texts belonging to its domain and (2) a phrase is more frequent than others in the same text. Based on this "competitive" conception of frequency, each phrase is assigned a score representing its significance, (not taking into account functional words). Phrases at the top of this ranking have the highest probability of being valid terms. This method can extract single-word terms as well as multi-word terms. On the other hand, it cannot locate terms which do not satisfy the statistical criteria, i.e. they are not frequent enough. This is partly due to the fact that it is difficult to draw the line between middle frequency and high frequency. Finally, the selected statistical formula can affect the performance of extraction in the same way that the selected rules of the grammar, i.e. its syntactical coverage, affect the performance of the grammatical method.

In between these methodologies stand other approaches which combine statistical processing with linguistic modeling [3], [4], [5], [6], [8], [9]. These hybrid systems initially construct a candidate term list using a term grammar and then filter this set through statistical techniques in order to remove syntactically acceptable phrases that are not "frequent" enough to be assigned valid termhood.

3. Method Description

The method we describe applies to Greek texts and the developed approach is largely based on the linguistic analysis of the source text. The output of the grammar is then statistically filtered. The basic steps are:

1. morphosyntactic annotation of the domain corpus using a transformation-based below part-of-speech tagger
2. parsing based on a pattern grammar endowed with regular expressions and feature-structure unification3
3. lemmatisation using a morphological lexicon and the grammatical category obtained by the tagger
4. statistical processing of the grammar-extracted terms, assigning a score to each potential term

The following flowchart illustrates how the algorithm proceeds:

3.1 Term Grammar

The pattern grammar4 used in the syntactic analysis is a subset of pattern rules presented in [10], whose rules cover a great part of the Greek terminology. It also utilises feature structure unification formalism (typical in grammar theories like HPSG) and regular expression operators. For example, the pattern that describes terms of the form: NOUN PREPOSITION NOUN has the following format:

Term pattern : (cat = Noun
\( \wedge (cat = Pronoun
\text{type} = Cl),\)
[ (cat = Prep
\text{type} = Sp);\n\wedge (cat = Art
\text{gender} = G
\text{number} = N
\text{case} = C) ] ;

1 For example, the typical rule Term ::- Adj+ Noun will definitely recognise many non-term phrases besides valid ones.
2 Assigning every noun with termhood creates a vast list of candidate terms, having a negative effect on precision, since few of them are real valid terms.
3 Unification is necessary in order to capture agreement between words (e.g. case agreement) in Greek.
4 We use the terms, pattern grammar and term grammar interchangeably
The '^' symbol at the end denotes optionality (zero or one appearance), the ';' symbol is the 'OR' operator and brackets are used to group elements. The basic constraint posed by this rule is the number-case-gender agreement between nouns and articles.

The term grammar consists of rules recognising two to four-word terms. Each rule was converted to a non-deterministic finite state automaton (NDFA). NDFA’s were used in preference to context-free grammar parsers (like Prolog DCG) because (a) they are much faster, operating in linear time (b) typical parsers do not support regular operators directly. Features used in unification include grammatical category as well as subcategorisation features like gender, case, tense, number, etc. Typical regular expression operators are optionality, kleene star, disjunction, etc.

Such a non-deterministic automaton is illustrated below:

\[
\begin{array}{c}
A \rightarrow \text{null} \\
\text{null} \rightarrow \text{Adj[Gen,Num,Case]\#(Adj[Gen,Num,Case]\#)*} \\
\text{Adj[Gen,Num,Case]} \rightarrow \text{Noun[Gen,Num,Case]} \\
\end{array}
\]

\( ? \) = optional, \( \# \) = zero or more times

3.2 Statistical filtering

After the term grammar module has been applied, the extracted terms are statistically evaluated in order to remove items without adequate statistical evidence. Daille’s extensive work [6] on contingency tables and statistical formulas used in valuing potential terms is a good source for such evidence. This filter was applied on two-word terms since manipulation of 2x2 contingency tables is easier. For a given pattern \( w_i \) \( w_j \), (e.g. noun+noun) the contingency table is defined as:

\[
\begin{array}{cc}
w_i & w_j \cup w_j' \\
w_i \cup w_j & a & b \\
w_i \cup w_j' & c & d \\
\end{array}
\]

\( a \) stands for the frequency of pairs involving both \( w_i \) and \( w_j \), (the number of occurrences of a pair)

\( b \) stands for the frequency of pairs involving \( w_i \) and \( w_j' \), (the number of occurrences of pairs where a given word appears as the first element of the pair)

\( c \) stands for the frequency of pairs involving \( w_i \) and \( w_j' \), (the number of occurrences of pairs where a given word appears as the second element of the pair)

\( d \) stands for the frequency of pairs involving \( w_i \) and \( w_j' \) (the total number of occurrences of the pairs where none of the given words appears as an element)

Daille [6] presents several statistical scoring formulas taken from lexical statistics, information retrieval and other technical domains like biology. All of these formulas were tested in a telecommunication corpus and the best scoring terms were examined against the Eurodicautom terminological bank (terminology data bank of the EEC, telecommunication section). The most effective formulas were found to be: (1) log-likelihood, (2) Fager and McGowan coefficient, (3) cubic association ratio, and (4) term frequency.

4. Method evaluation

The corpus on which the system was tested is a software manual of about 90000 words in size. The particular text was selected because it also contained a human-crafted terminology index against which the results would be evaluated. During the evaluation procedure terms are reduced to their canonical form where each word is replaced by its lemma.

3 By words, in this context, we refer to content words.
along with the manual. Since capturing single-word terms was not within the scope of our experiment with the term grammar, the target list was first cleared from single-word terms, resulting in 209 target terms in total. It is worth noting here, that some terms in the index were not exactly (or not at all) matched in the text due to slight variations (or total absence from the text!), resulting in a smaller number of target terms, 204. The grammar extracted 3596 candidate terms, 130 of them being correct, i.e. matching correctly one of the target 204 ones. Thus recall and precision were calculated as follows:

Recall: \[ \frac{130}{204} = 63.7\% \]

Precision: \[ \frac{130}{3596} = 3.6\% \]

The recall figure is considered satisfactory. On the other hand, the precision figure is considerably, but not unexpectedly, low since grammars inherently produce numerous terms because of the generality of the rules.

Study of the terms that were not recognised revealed that 17% of them were due to foreign words or to tagger inaccuracies caused by unknown words or incorrect disambiguation. The latter can be either incorrect post identification (e.g. adverb instead of adjective or verb instead of noun) or disagreement in below post features (e.g. assignment of different case in two successive words, participating in a term, for instance: adjective and noun).

Regarding the evaluation of the statistical processing module, we conducted the same experiment, using these scores (adding another one, NC-value [9]), and found that the best results came from the following four formulas:

1. **Fager and McGowan Coefficient (FAG)**
   \[ \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2}\sqrt{a+b} \]

2. **Cubic Association ratio (IM3)**
   \[ \log_2 \frac{a^3}{(a+b)(a+c)} \]

3. **Log-likelihood (LLH)**
   \[ a \log a + b \log b + c \log c + d \log d - (a+b) \log(a+b) - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) + (a+b+c+d) \log(a+b+c+d) \]

4. **NC-value**
   a formula which takes into account both the number of appearances of the term in the corpus and the context information, i.e. verbs, adjectives and nouns that surround the evaluated term.

These results are very close to the top-ranked formulas found in Daille's work. The following table summarises precision and recall figures of the top 200 terms produced by each one of the four different scorings. Since the grammar located 77 out of 134 two-word terms in the terminology index we provide two recall figures, with two different denominators: the first referring to the full index (134), the second referring to the grammar results (77).

<table>
<thead>
<tr>
<th></th>
<th>FAG</th>
<th>IM3</th>
<th>LLH</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>17%</td>
<td>14%</td>
<td>18%</td>
<td>20%</td>
</tr>
<tr>
<td>Recall</td>
<td>26%</td>
<td>45%</td>
<td>22%</td>
<td>38%</td>
</tr>
<tr>
<td></td>
<td>28%</td>
<td>48%</td>
<td>30%</td>
<td>52%</td>
</tr>
</tbody>
</table>

A critical issue when evaluating statistically produced term lists is the cut-off threshold, i.e. the number of top-scoring terms we decide to evaluate. In our experiment the selected threshold is 200, since the number of target terms is 204. Generally, this number should not be very low (low recall) or high (low precision) and is also affected by the domain and the user's requirements.

5. **Conclusion and outlook to future work**

We have presented a system for automatic term extraction using both linguistic and statistical modelling. Linguistic processing is performed through an augmented term grammar, the results of which are statistically filtered using frequency-based scores. The method was able to locate 62% of technical terminology in a software manual text, compared against a hand-crafted terminology index of it.

The presented work is currently enriched with features aiming at improving its efficiency by reducing the number of potential terms recognised by the grammar module and improving its coverage. These features include:

- Implementation of other statistical processing techniques (such as term weighting with TF-IDF scoring [13]) of the terms extracted by the grammar so that the frequency of the term in a reference corpus of the domain is also taken into consideration. Furthermore, TF-IDF can also score and locate candidate single-word terms.

---

6 Contingency tables utilise only the frequency of the word inside the processed text.
• Utilisation of further syntactic information (NP head) in order to group together terms with the same semantic but slightly different syntactic structure

• Extension of an already existing terminological base through linguistic operations such as overcomposition, modification, coordination, etc.

6. References


