A variational linguistic approach to term extraction.

Dirk De Hertog¹, Kris Heylen¹, Dirk Speelman¹, Hendrik Kockaert¹,²

¹QLVL
University of Leuven
Blijde-Inkomststraat 21
3000 Leuven (Belgium)

²Department of Applied Language Studies
Lessius University College
Sint-Andriesstraat 2
2000 Antwerpen, Belgium

{dirk.dehertog, kris.heylen, dirk.speelman}@arts.kuleuven.be, hendrik.kockaert@lessius.eu

Abstract

This paper describes how a toolset developed within variational linguistics for the purposes of identifying regional variants, can be used in the field of term extraction. The notion of stable lexical marker will be introduced as a method to quantify termhood as a function of both high relative frequency and uniform dispersion of words in a specialised domain. As such, the work is an extension of so called contrastive approaches to term extraction. The Belgian legal domain will serve as a case study.

1. Introduction

In the field of term extraction there have been many approaches that focus on the extraction of multiword units, assuming that multiword units comprise the majority of terms in most subject fields. Lately some studies have emerged that stress the importance of simple or monoword extraction. Not only is the prevalence of multiword terms an insufficient reason to disregard monoword terms, the exact ratio of mono- to multiword terms is hard to verify and might be domain dependent. While methods that extract multiword terms can function reasonably well using domain-internal frequency data, this is not sufficient to distinguish monoword terms from non-terms. Therefore, monoword term extraction has to rely on contrasting domain-internal with domain-external corpus evidence. These are so-called constrastive term extraction methods.
2. Contrastive approaches

2.1. State-of-the-art

Contrastive approaches rely on the fact that terms occur more frequently in their domain than they do in a general corpus. Several researchers have been using such a contrastive approach to determine termhood, the degree to which a word is considered to be a term. Tfidf is the oldest contrastive measure in use, measuring the word's termhood as a combination of its frequency and the inverted document frequency. Gillam and Ahmad (2005) use a measure they refer to as the weirdness of a word, which is defined as the result of the comparison of the word’s normalised frequencies between a specialised corpus and a general language corpus. In this manner they "identify signatures of a specialism". Those words which combine high frequency and high weirdness are of most interest for term identification. Kit and Liu (2008) quantify the termhood of a term candidate as its difference in frequency rank between a domain and a background corpus. This rank is based on the word's frequency for both types of corpora and is normalised by the total number of types in the corpus' vocabulary. Chung (2003) uses a normalised frequency ratio to decide on termhood. Wong (2007) proposes a similar technique that uses distributional behaviour of a word in opposing corpora to measure what he calls intra-domain distribution and cross-domain distributional behaviour. The first distribution is used to calculate a domain prevalence score, which measures the extent of the term's usage within the target domain. The second distribution is the basis for a domain tendency score, which measures the extent of term usage towards the target domain. Drouin (2008) compares precision and recall for the ranking of different hypothesis testing methods, trying to determine which method works best. The method proposed in this paper is also a contrastive approach but with its origin in variational linguistics.

2.2. A variational linguistic perspective

The method we propose to use for term extraction was originally developed in the cross section between corpus linguistics (Kilgariff, 2001) and variational linguistics in the Labovian tradition. Speelman, Gondelaers and Geeraerts
(2006) implemented a toolset to identify so-called *lexical markers* of different language varieties. An example is the difference in word use for the concept **UNDERGROUND TRANSPORTATION NETWORK** in American and British English. In this case, *subway* is said to be a lexical marker for American English and *underground* for British English. More specifically, the tool relies on statistical hypothesis testing of which words occur significantly more frequently in a corpus that is representative of a given variety when compared with a corpus from another language variety. Important in the current context is that the specialised domains studied in terminology research can be considered as a specific language variety that is different from general language. The method developed for identifying lexical differences in two varieties, can thus be used for identifying terms using a specialised and a general corpus. Being a lexical marker for a specialised corpus, can be seen as one of the necessary characteristics for qualifying as a term. Interestingly, the tool does not limit itself to a straightforward comparison of relative frequencies. It offers an additional functionality which allows to measure a word's dispersion in a variety-specific corpus. To our knowledge, this technique has not yet been applied in terminology research, but could potentially benefit term-extraction from domain-specific corpora.

### 2.3. Stable lexical marker analysis

Stable lexical marker analysis defines the dispersion of a word as its consistency and stability within the domain and calculates this by using a pairwise comparison of a subdivision of both the reference corpus and the analysis corpus. For example, both the specialized corpus (S) and the reference corpus (R) might be divided into 8 parts: \{S₁, S₂, ..., S₈\} and \{R₁,R₂,..., R₈\}. The next step is a pairwise comparison between all of the S-members and all of the R-members: \{S₁, R₁\}, \{S₁, R₂\}, ... \{S₈, R₈\}. In each pairwise comparison, statistical hypothesis testing determines which words are *lexical markers* that occur significantly more frequently in the specialized corpus as compared to the reference corpus. A scoring scheme is applied so that a word gets credit for each pairwise comparison in which it is a lexical marker. If a word obtains a high score over all pairwise comparisons, it is called a *stable* lexical marker. For the example above,
there are 64 possible combinations between group S and group R so the maximum score is 64 and the minimum score is 0. This way, the analysis provides a ranking that assigns the highest scores to the words that most consistently occur with a significantly higher frequency in the specialised corpus as compared to the general corpus. In sum, the lexical stable marker method takes into account two properties of variety-specific lexical items. As other contrastive approaches, it extracts words that have an above-expected frequency in the specialised corpus, but additionally, the method assures that these words have a high dispersion in the specialised corpus. This has the advantage of filtering out any frequency bias that might be introduced by just a part of the corpus.

3. Case study: the Belgian legal domain

The Stable Lexical Marker method will be applied to mono-word term extraction from Dutch texts in the Belgian legal domain. Although the variational linguistic notion of stable lexical marker of a language variety does not completely overlap with the notion of a term in terminology research, the method is well-suited for the legal domain, as it is characterised by a very specific linguistic style that goes beyond the presence of terminological units in a strict sense, but also involves rhetorical expressions and idiomatic phrases. When these are considered as terminologically relevant LSP-characteristics - and we think they should - a linguistically underpinned analysis method like stable lexical marker analysis, might be better suited for terminology extraction than traditional term extraction methods.

As a specialised corpus we have at our disposal a bilingual translation corpus obtained from one of the partners of the project at hand, the Belgian federal Justice department. This corpus contains about twelve million Dutch words. As a Dutch reference corpus we have newspaper material collected over several years totalling over one billion words. A subset of this material will be selected.

In a first step the corpora under investigation undergo linguistic preprocessing, such as lemmatizing and part-of-speech-tagging. In a second step the corpora are subdivided in different random samples. The size of the available corpora makes
it possible to maintain a high enough frequency for the analysis of salient words. Frequency is obtained for all lemma's in the two subcorpora, along with the frequency of each bin. Because the comparison method is statistical in nature, low frequency words from the specialised corpus are filtered out. Special attention is given to the effect of the subdivision on the increase of low-frequency words and how trustworthy the results for the general analysis for these words was in the first place. Both a complete comparison and a pairwise comparison as described above will be carried out to identify lexical markers and thus measure their termhood.

4. Evaluation

As a gold standard, a term base offered by the Belgian federal justice department will be used. Precision and recall will be calculated based on this term base for both the general contrastive method, and the stable lexical marker analysis method. The method will also be benchmarked against a subset of the contrastive term extraction methods outlined in section 2.

References


