Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure

Delphine Bernhard
TIMC-IMAG
Institut de l'Ingénierie et de l'Information de Santé
Faculté de Médecine
F-38706 LA TRONCHE cedex
Delphine.Bernhard@imag.fr

Abstract

Morphologically complex terms composed from Greek or Latin elements are frequent in scientific and technical texts. Word forming units are thus relevant cues for the identification of terms in domain-specific texts. This article describes a method for the automatic extraction of terms relying on the detection of classical prefixes and word-initial combining forms. Word-forming units are identified using a regular expression. The system then extracts terms by selecting words which either begin or coalesce with these elements. Next, terms are grouped in families which are displayed as a weighted list in HTML format.

1 Introduction

Many methods for the automatic extraction of terms make use of patterns describing the structure of terms. This approach is especially helpful for multi-word terms. Depending on the method, patterns rely on morpho-syntactic properties (Daille, 1996; Ibekwe-SanJuan, 1998), the co-occurrence of terms and connectors (Enguehard, 1992; Baroni and Bernardini, 2004) or the alternation of informative and non-informative words (Vergne, 2005). These patterns use words as basic units and thus apply to multi-word terms. Methods for the acquisition of single-word terms generally depend on frequency-related information. For instance, the frequency of occurrence of a word in a domain-specific corpus can be compared with its frequency of occurrence in a reference corpus (Rayson and Garside, 2000; Baroni and Bernardini, 2004). Technical words usually have a high relative frequency difference between the domain-specific corpus and the reference corpus.

In this paper, we present a pattern-based technique to extract single-word terms. In technical and scientific domains like medicine many terms are derivatives or neoclassical compounds (Cottez, 1984). There are several types of classical word-forming units: prefixes (extra-, anti-), initial combining forms (hydro-, pharmaco-), suffixes (-ism) and final combining forms (-graphy, -logy). Interestingly, these units are rather constant in many European languages (Namer, 2005). Consequently, instead of relying on a subword dictionary to analyse compounds like (Schulz et al., 2002), our method makes use of these regularities to automatically extract prefixes and initial combining forms from corpora. The system then identifies terms by selecting words which either begin or coalesce with these units. Moreover, forming elements are used to group terms in morphological and hence semantic families. The different stages of the process are detailed in section 2. Section 3 describes the results of experiments performed on four corpora, in English and in French.

2 Description of the method

2.1 Extraction of words

The system takes as input a corpus of texts. Paragraphs written in another language than the target language are filtered out. Texts are then tokenised and words are converted to lowercase. Besides, words containing digits or other non-word characters are eliminated. However, hyphenated words are kept since hyphens mark morpheme boundaries. This preliminary step produces a word frequency list for the corpus.

2.2 Acquisition of combining forms

Prefixes and initial combining forms are automatically acquired using the following regular expression: **([aio]-)?(\w{3,}[aio])-**. This regular expression represents character strings whose length is higher or equal to 4, ending with *a*, *i* or *o* and immediately followed by a hyphen. The first part of the regular expression accounts for words where several prefixes or combining forms follow one another (as for instance in the French word "**hépato-gastro-**entérologues"). This regular expression applies to English but also to other languages like French or German: see for instance "**chimio-**radiothérapie" in French, "**chemo-**radiotherapy" in English or "**Chemo-**radiotherapie" in German.

2.3 Identification of terms

Terms are identified using the following pattern describing their morphological structure: **E+W** where **E** is a prefix or combining form and **W** is a word whose length is higher than 3; the '+' character represents the possible succession of several **E** elements at the beginning of a term. Prefixes and combining forms may be separated by a hyphen. When this pattern applies to one of the words in the corpus, two terms are recognised, one with a **E+W** structure and the other with a **W** structure. For instance, given the word "ferrobasalts", the system identifies the terms "ferrobasalts" (E+W) and "basalts" (W).

2.4 Conflation of terms

Term variants are grouped in order to ease the analysis of results. The method for terms conflation can be decomposed in two stages:

1. Terms containing the same word W belong to the same family, represented by the word W. For instance, both "chemotherapy" and "radiotherapy" contain the word "therapy": they belong to the same family of terms, represented by the word "therapy".

2. Two families are merged if they are represented by words sharing the same initial substring (with a minimum initial substring length of 4) and if the same prefix or combining form occurs in one term of each family. Consider for instance the families F_1= [oncology, psycho-oncology, radio-oncology, neuro-oncology, psychooncology, neurooncology] and F_2 = [oncologist, neuro-oncologist]. The terms representing F_1 ("oncology") and F_2 ("oncologist") share an initial substring of length 7. Moreover the terms "neuro-oncology" from F_1 and "neuro-oncologist" from F_2 contain the combining form "neuro". Families F_1 and F_2 are therefore united.

When terms have been conflated, we select the most frequent term as a family's representative.

2.5 Data visualisation

The results obtained are displayed as a weighted list in HTML format. Such lists, also named "heat maps" or "tag clouds" when they describe tags[1] usually represent the terms and topics which appear most frequently on websites or RSS feeds (Wikipedia, 2006). They can also be used to represent any kind of word list (Véronis, 2005). Different colours and font sizes are used depending on the word's frequency of occurrence. We have adapted this method to visualise the list of extracted terms. Since several hundred terms may be extracted, only the terms representing a family are displayed on the weighted list. Weight is given by the cumulated frequency of all the terms belonging to the family (see Figure 1).

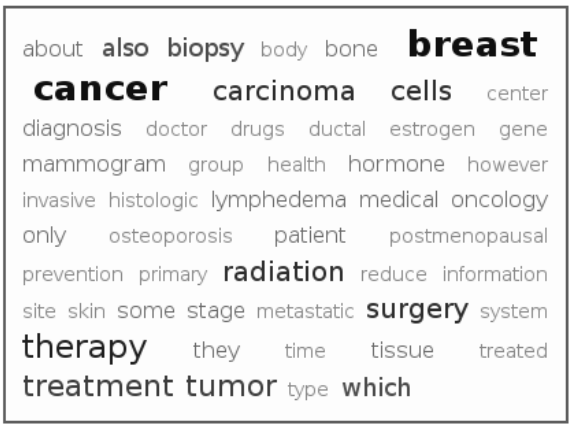
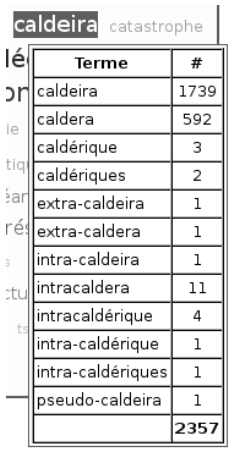


Figure 1: Term cloud example (Corpus: BC_en)

Further information (terms and frequencies) is displayed thanks to tooltips (see Figure 2), using the JavaScript overLIB libray (`http://www.bosrup.com/web/overlib`).

[1] See for example TagCloud: `http://www.tagcloud.com`



Terme	#
caldeira	1739
caldera	592
caldérique	3
caldériques	2
extra-caldeira	1
extra-caldera	1
intra-caldeira	1
intracaldera	11
intracaldérique	4
intra-caldérique	1
intra-caldériques	1
pseudo-caldeira	1
	2357

Figure 2: Detailed term family displayed as a tooltip (Corpus: V_fr)

3 Experiments and results

3.1 Corpora

The system has been experimented on 4 corpora covering the domains of volcanology (V) and breast cancer (BC), in English (en) and in French (fr). The corpora have been automatically built from the web, using the methodology described in (Baroni and Bernardini, 2004), via the Yahoo! Search Web Services (`http://developer.yahoo.net/search/`). The size of the corpora obtained are given in Table 1. This table also gives the number of key words, i.e., single-word terms extracted by comparing the frequency of occurrence of words in both corpora for each language (Rayson and Garside, 2000). Only terms with a log-likelihood of 3.8 or higher ($p<0.05$) have been kept in the key words list. Table 2 gives a numerical overview of the results obtained by our method.

Corpus	Tokens	Word forms	Key words
BC_fr	1,451,809	46,834	13,700
BC_en	7,044,146	88,726	17,602
V_fr	1,777,030	59,909	13,673
V_en	2,929,591	48,257	19,641

Table 1: Size of the corpora

3.2 Prefixes and initial combining forms

As shown by Table 2, the number of prefixes and initial combining forms identified is proportionally less for the volcanology corpora both in English and in French. Medical corpora seem to be more adapted to the method since the number of terms extracted is higher. The prefixes and combining forms identified are also highly dependent on the corpus domain. For instance, amongst the most frequent combining forms extracted for the BC corpora, we find “radio” and “chemo” (“chimio” in French) and for the V corpora, “strato” and “volcano”.

Corpus	Word-forming elements	Terms	Term families
BC_fr	334	4,248	911
BC_en	382	5,444	1,338
V_fr	182	1,842	583
V_en	188	1,648	564

Table 2: Number of word-forming elements, terms and term families identified for each corpus

3.3 Terms

The overlap percentage between the list of terms and the list of key words ranges from 38.65% (V_fr) to 56.92% (V_en) of the total amount of terms extracted. If we compare both the list of key words and the list of terms extracted for the BC_en corpus with the Unified Medical Language System Metathesaurus (`http://www.nlm.nih.gov/research/umls/`) we notice that some highly specific terms like “disease”, “blood” or “x-ray” are not identified by our method, while they occur in the key words list. These are usually morphologically simple terms, also used in everyday language. Conversely, terms with low frequency like “adenoacanthoma”, “chondroma” or “mammotomy” are correctly identified by the pattern-based approach but are missing in the key words list. Both methods are therefore complementary.

In some cases, stop-words are extracted. This is a side effect of the pattern used to retrieve terms. Remember that terms are words which coalesce with combining forms, possibly with hyphenation. In English hyphens are sometimes mistakenly used instead of the dash to mark comment clauses. Consider for instance the following sentence: “*As this* **magma-which** *drives one of the worlds largest volcanic systems-rises, it pushes up the Earths crust beneath the Yellowstone Plateau.*”. Here “magma” is identified as a combining form since it ends with ‘a’ and is directly followed by a hyphen. Consequently, “which” is wrongly identified as a term.

3.4 Term families

Several types of term variants are grouped by the term conflation algorithm: (a) graphical and orthographical variants like "tumour" (British variant) and "tumor" (American variant); (b) inflectional variants like "tumor" and "tumors"; (c) derivational variants like "tumor" and "tumoral".

Two types of conflation errors may however occur: over-conflation, i.e., the conflation of terms which do not belong to the same morphological family and under-conflation, i.e. the absence of conflation for morphologically related terms. Some cases of over-conflation are obvious, such as the grouping of "significant" with "cant". In some other cases it is more difficult to tell. This especially applies to the conflation of terms composed of word final combining forms like "-gram" or "-graph". Under-conflation occurs when no combining form is shared between terms belonging to families represented by graphically similar terms. For instance, the following term families are extracted from the French volcanology corpus (V_fr): F_1= [basalte, métabasalte, méta-basalte], F_2= [basaltes, ferro-basaltes, paléobasaltes] and F_3= [basaltique, andésitico-basaltique]. These families are not conflated, even though they obviously belong to the same morphological family.

4 Conclusion

We have presented a method for the automatic acquisition of terms from domain-specific texts using morphological structure. The method also groups terms in morphological families. Families are displayed as a weighted list, thus giving an instant overview of the main topics in the corpus under study. Results obtained from the first experiments confirm the usefulness of a morphological pattern based approach for the extraction of terms from domain-specific corpora and especially medical texts. The method for the identification of compound words could be improved by an automatic approach to morphological segmentation as done by (Creutz and Lagus, 2004). Term clustering could be ameliorated as well by investigating the usefulness of stemming to avoid under-conflation.

References

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1313–1316.

Henri Cottez. 1984. *Dictionnaire des structures du vocabulaire savant. Éléments et modèles de formation.* Le Robert, Paris, 3rd edition.

Mathias Creutz and Krista Lagus. 2004. Induction of a Simple Morphology for Highly-Inflecting Languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51.

Béatrice Daille. 1996. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, Massachusetts.

Chantal Enguehard. 1992. *ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique*. Ph.D. thesis, Université de Technologie de Compiègne.

Fidelia Ibekwe-SanJuan. 1998. Terminological variation, a means of identifying research topics from texts. In *Proceedings of the Joint International Conference on Computational Linguistics (COLING-ACL'98)*, pages 564–570.

Fiammetta Namer. 2005. Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. In *Actes de TALN 2005*, pages 63–72.

Paul Rayson and Roger Garside. 2000. Comparing Corpora using Frequency Profiling. In *Proceedings of the ACL Workshop on Comparing Corpora*, pages 1–6.

Stefan Schulz, Martin Honeck, and Udo Hahn. 2002. Biomedical Text Retrieval in Languages with a Complex Morphology. In *Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain*, pages 61–68.

Jacques Vergne. 2005. Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. In *Actes de la Conférence Internationale sur le Document Électronique (CIDE 8)*, pages 155–168.

Jean Véronis. 2005. Nuage de mots d'aujourd'hui. `http://aixtal.blogspot.com/2005/07/lexique-nuage-de-mots-daujourdhui.html`. [Online; accessed 31-January-2006].

Wikipedia. 2006. RSS (file format) — Wikipedia, The Free Encyclopedia. `http://en.wikipedia.org/w/index.php?title=RSS_(file_format)&oldid=37472136`. [Online; accessed 31-January-2006].