Use of Greek and Latin forms for term detection

R. Estopà; J. Vivaldi; M. T. Cabré

rosa.estopa@trad.upf.es jorge.vivaldi@info.upf.es teresa.cabre@trad.upf.es
Institute for Applied Linguistics
Universitat Pompeu Fabra
Rambla Santa Mònica, 30
08002 Barcelona, Spain

Abstract

It is well known that many languages make use of neo-classical compounds, and that some domains with a very long tradition like medicine made an intense use of such morphemes. This phenomenon has been largely studied for different languages with the common result that a relatively short number of morphemes allows the detection of a high number of specialised terms to be produced. We believe that the use of such morphological knowledge may help a term detector in discovering very specialised terms.
In this paper we propose a module to be included in a term extractor devoted specifically to detect terms that include neo-classical compounds. We describe such module as well the results obtained from it.

1. Introduction

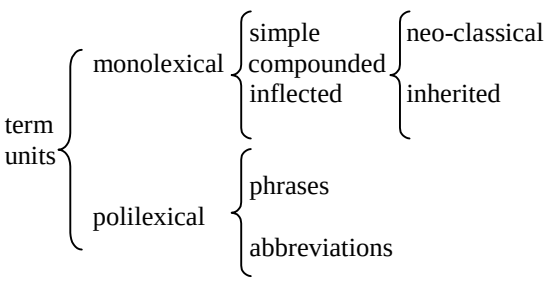
Most of the known automatic term detection systems use term dictionaries or simple morphosyntactic patterns typically used to represent complex terms. Some authors (Cabré et. al., in press, Kageura, 1996) have pointed out that although this strategy detects many of the terms that appear in specialised texts it has some drawbacks:

- monolexical units are not detected
- a lot of noise is produced, that is many term candidates are not pertinent to the domain although their syntactical patterns are correct
- some of the actual terms are not detected by extractors. This problem of the silence appears in statistical or hybrid systems

Taking into consideration the above points we consider that for improving the efficiency of term extractors it is necessary to use different kinds of linguistic resources: morphological, syntactic, semantic and pragmatic in other to improve the current precision and recall figures.

2. Morphological features of terms

The basic morphological structure of terminological units (TU) is roughly the same as units from general language. Thus, from a structural point of view terms, like words, can be monolexical or polilexical. According to the morpheme number and morpheme type, there can be simple, inflected, compounded, shortened and phrasal TU. Thus in any specialised domain, terms of the following nature can be found:



However it its true that in all specialised domains these types are to a greater o lesser extend represented, each domain selects a particular number of structures and type of units to designate its specialised units whereas other are less represented. Thus for instance in medicine —as we shall see later— there are much more neo-classical compounds than in technological domains such as computer science, mechanics or cinematography. Furthermore, if we consider specifically inflected or compounded terms we realise that, given a particular domain, there is a regular appearance of specific forms. It can be seen, for example, in technology where the suffix *-atge* is mostly used to refer to technical operations (*blindatge* 'shield'; *cribatge* 'sieving'; *rodatge* 'shooting'; etc.). In contrast, in biomedicine most of nominalisations are formed through the suffix *-ció* (*inflamació* 'inflammation'; *tumefacció* 'tumefaction'), being *-atge* set aside. This regularity can also be found within categories. Thus, in the medical pathology class the neo-classical form *-itis* is almost compulsorily used to refer to inflammations (*hepatitis*, *arthritis*, *mastittis*, *meningitis*, etc.). The same holds for *-oma* with regard to tumours (*carcicoma*, *hepatoma*, etc.).

Taking into account these morphopragmatic features of terms, we state that models of designation in a specialised domain are grounded in the systematic selection particular properties and features. Thus terminology extractors could make use of these morphological regularities of terms in order to ameliorate their results.

3. Greek and Latin forms

It is well known that many languages make use neo-classical compounds, and that some domains with a very long tradition such as medicine, biology and all disciplines related to the health sciences make use of such morphemes. This phenomenon has been largely studied with regard to the English language (Smith et al., 1996) as well as Spanish (Lopez et al, 1990) and Catalan (Bernabeu et al, 1995) among others. The common finding is that a relatively short number of Greek and Latin forms (stems,

prefixes and affixes) yield to a high number of specialised terms.

We have shown that between 30% and 40% of the monolexical terminological units found in medical texts both in Catalan and Spanish include a Greek and Latin form (*biopsy*, *meningitis*, *nephritis*, *carcicoma*, *rhinoplatia*, etc.). Further, the vast majority of these units are the head of polilexical terminological units (*endoscopic biopsy*, *cerebrospinal linfocitic meningitis*; *intersticial nephritis*, *bronchogenic carcicoma*, *Joseph rhinoplasty*, etc.) (Estopà, 1999).

Usually this knowledge is used by health specialists to recognise and understand most of the specialised terms of their domain and to build new term units following a particular conceptual paradigm.

To illustrate this let us consider the term meninge. It is builded from the Greek stem mênigx-, méniggos which means membrane, and it is used to designate any of the three membranes that surround the encephalon and the spinal marrow. A great deal of units are formed from the stem mening-o. Meningitis, for instance, is composed of a stem (mening) and a suffix (-itis, meaning 'inflammation of a body part'), being the overall meaning achieved compositionally. This term can be recognised in two ways: either by adding it to the lexicon or by reconstructing its formation (mening + itis). Likewise this stem is also used to form many other related terms (meningioma, meningocele, meningomyelitis, leptomeninges, meningocerebritis, meningorrhea, meningocephalopathy, etc). Often mening and other similar stems are part of hybrid units, that is, they appear in combination with morphemes and lexemes from the general language (meningitic; meningorecurrence). Also they may occur in complex units (otitic meningitis, aspeptic meningitis, external meningitis, Quicke's meningitis, etc.) in which three level of analysis are observed: forms, monolexical TU and polilexical TU. It should be emphasised that, unlike what happens in the second and third morphological level, units from the first level cannot be regarded as TU but neo-classical forms. However, only by controlling units from the first level can it be possible to recognise and integrate units from the two other groups.

Accordig to what have just been exposed, we believe that, as far as health sciences is concerned, term detection systems may benefit from a particular module so as to recognise terms grounded in Latin and Greek forms. This will lead to obtain better results without having to tag and/or add all candidate words in the lexicon, which is costly, unreliable and time-consuming.

4. Term extraction

To build an efficient automatic extractor of candidate terms we assume the following two statement:

Each type of word (simple, inflected and compounded) has its own specific linguistic features and, accordingly, automatic extraction techniques have to benefit from them.

Taking into account that there are linguistic differences, we should avoid using a single technique for all MTU, even for some types of them. As a result, we state that technique combination is the most useful strategy to follow in order to benefit from them.

According to the above statements, each term extractor should contain several modules adapted to the features of each linguistic characteristic (Cabré *et al.*, 2000). This is consistent with the fact that the results of a term extractor combining knowledge from several independent classifiers are likely to be improved (Vivaldi & Rodríguez, 2000). We are applying these ideas in the building of a term extractor for medical terms which comprises several modules, being one of them specifically designed to benefit from terms grounded in neo-classical compounds. This extractor includes other modules: semantic content extractor, context analysis and statistical technique.

4.1. Neo-classical form module

Although the neo-classical form module may function as a morphological analyser recognising words from running text its main goal is to recognise only those words composed by Greek and Latin forms with a reduced set of affixes. Whenever possible the information provided consists in the decomposition of the analysed words into its components. The overall meaning of words can be restored compositionally from the meaning of each of their components.

To attain the above goals the module is provided with a dictionary containing all particular stems, affixes and their related information. Thus it comprises a whole word and its hyperonym. All this information is combined with a set of rules so as to control the combination of stem, affixes and auxiliary words which altogether lead to a finite state automata.

Word analysis consists in a sequence of transitions from a state labelled "START" to a final one labelled "END". For instance, let us take the word hepatitis. It can be analysed as two-sequence transition: first, from "START" to an intermediate node 1 involving hepat, and second, from node 1 to "END" involving itis. Hepat is a stem being associated the word 'liver' and its hyperonym 'digestive apparatus". Meanwhile itis is a suffix being associated the word 'inflammation'. Figure 1 shows these two transitions yielding to the recognition of the word hepatitis.

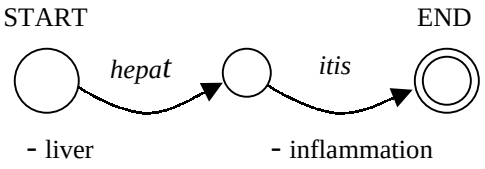


Figure 1 Transitions for *hepatitis*

Similarly many complex other complex word can be recognised. Figure 2 shows part of the transitions that leads to the recognition complex medical words like *electrocardiogram, histology, intravenous, bronchopneumonia*, etc.:

This mechanism has many advantages against traditional ones, especially in relation to time and space reduction.

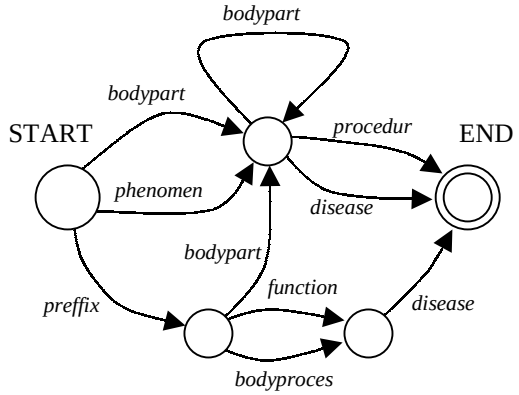


Figure 2 Sample of all transitions

This mechanism that is being proposed takes into account the different value of Greek and Latin forms. Thus, if a candidate term has a suffix or a stem with a medical sense, then it will be proposed by the tool as terminological although it could not be splitted into all its elements[1]. Thus we have observed that in a medical texts lexical units containing suffixes such as *–itis*, *-osis*, *-oma*, *-pathy*, *-scope*, *-lysi*, etc. and stems such as *trombo-*, *cardio-*, *aorto-*, *pneumo-*, *rhino-*, etc. will hardly acquire a terminological valule. The condition posed above allow to detect neo-classical miexed compounds formed by a neo-classical form and a general-language word (e.g. *bronchodilatador*, *pharmacogenetics*).

In contrast, other different group of forms, especially prefixes and, to a lesser extend, a number of suffixes do not show a medical sense. These are formants of a more general meaning. Further, their meaning is transversal inasmuch as they are used to form terms in many other specialised domains as well as general-language words. Broadly speaking theses prefixes indicate location (*ambi*, *inter- hypo-*, *sub*, etc.), number (*bi-*, *tetra-*, *poly-*, *mono-*, etc.), colour (*leuco-*, *melano-*, *cloro-*, *rubeo-*, etc.), measure (*macro-*, *micro-*, *hypo-*, *hyper*, etc.), quality (*tachy-*, *homo-*, *neo-*, *pseudo-*, *auto-*, etc.) and direction (*dia-*, *circum-*, *ultra-*, etc.). In all these cases it should be avoided the overgeneration of false candidates so the tool proposes as term candidates those units which can be decomposed into all their elements.

Let us take as an example the analysis obtained by our module regarding the Spanish term *cardiopatía "cardiopathy"*:

Component	Type	Reference to
cardi	stem	Heart (cardiovascular apparatus)
o	link vowel	-
pat	stem	Disease (diseases)
ia	suffix	Pathological state

We validate the specialised nature of this term analysing the "reference to" components (*heart* and *disease* in this example) against the lexical database EuroWordNet[2]. This simple mechanism allows the detection of a considerably large number specialised units, mainly monolexical, but also polilexical, which otherwise would be very difficult, or impossible, to detect. Further examples of the analysis obtained are the following:

Components	Type	Reference to
• ***Linfadenopatia***		
linf	Stem	linfa (cardiovascular apparatus)
adeno	Stem	ganglio (cardiovascular apparatus)
pat	Stem	disease (diseases)
ia	Suffix	pathological state
• ***Linfocitosis***		
linf	Stem	lymph (cardiovascular apparatus)
o	link vowel	-
cit	Stem	cell (tissue)
osis	Suffix	pathological state
• **Microhematuria**		
micro	Preffix	-
hemat	Stem	blood (humores)
ur	Stem	urine (humores)
ia	Suffix	pathological state
• ***Trombocitopenia***		
tromb	Stem	coagulum (disease)
o	link vowel	-
cit	Stem	cell (tissue)
o	link vowel	-
pen	Suffix	scarce (quantity)
ia	Suffix	pathological state

It should be noted that not all modules of our extractor being developed show the same value for all types of terms. Hences as for neo-classical compounds, which this paper's main focus, the most important module is the neo-classical form module and lexical database module and context module are complementary. By contrast, to detect simple terms the neo-classical module is not pertinent, being the main one the lexical database module and the context modules remains complementary.

5. Experiment test

The above technique has been applied to a highly specialised medical corpus on Rickettsial diseases from Farreras-Rozman's *Medicina interna* —Internal Medicine—(1997). This firs experiment test has been restricted to monolexical terms and further research will be devoted to polilexical terms whose head or adjunct is a neo-classical compound.

A term candidate extractor based in Bourigault (1994), proposes 645 units as being MTU candidates of 12.069 occurrences analysed by our module. To evaluate the performance of automatic techniques we have considered the manual terminological retrieval made by three specialists in the domain.

[1] This behaviour can be configured.

[2] EWN is a general purpose multigual lexical DB based on Princeton WordNet covering Spanish and other European languages. Wordnets are structured in lexical-semantics units (synsets containing a set of synonymous words) linked themselves with basic semantic relations. See (Vossen, 1999) for details.

6. Result analysis

From all the proposed MTUs selected by specialists (312), 114 (33.6%) include neo-classical compounds. The results of this selection and the application of all the extraction modules from our system are as follows:

Detection method	Numberof MTUs	
	%	#
Manual	-	114
- neo-classical compounds	87.7	100
- context analysis	65	74
- semantic contents extractor	35	40

The above table shows the MTU detection results comprising neo-classical compounds as well as the remaining modules included in our term extractor for MTU detection. The recall figure is of 87.7 %, which refers to all MTUs containing neo-classical compounds. It represents a 32.5 % in relation to the whole number of the MTUs within the text. The fact that some terms are also detected by other mechanism, mainly by the semantic contents extractor, indicates that some neo-classical compounds are very common (*bronchitis*, *embolism*, *arthrosis*, etc.). It should also be noted the relevance of this module because otherwise a certain number of candidates could not be detected by the other methods.

The strategy to accept candidates only having a medical stem among its components has shown to be successful in our specialised text because it permits to detect high specialised terms such as *hepatomegalia* and *meningococcemia.*

No matter how good these results are, detection of terms formed by neo-classical forms is still fraught with problems:

This tool accept units including part that formally coincide with a particular neo-classical. This holds for *embarazada* (pregnant woman) which is decomposed by the module into *em*[3] *(blood), bar (pressure)* and *azada* (part that cannot be analysed). Since a neo-classical compound has been encountered, then it is proposed as term candidate.

Here this unit can be taken as term in this text but the meaning attributed by the extractor is unreliable as it does not contain any medical term. Rather, it has been formed through other word forming mechanisms. To solve this problem, as sort of stronger restriction should be applied to neo-classical compounds and meanwhile giving a more important role to the linking vowel with the neo-classical compound (*o* or *i*).

A second point has do to with neo-classical prefixes and some suffixes with no medical sense that are combined with words having medical sense. Thus, for instance the *mono-* prefix allows term formation provided it is combined with a regular words such as *monodose* = mono + dose. However, it can also find *monochrome* which contains the *mono-* prefix without being a medical term. As for a general sense suffix we can consider words such us *radiograph*, *ecograph*, *mamograph* detected by some preliminary tests. These words turn out to be terms in our text. However, there is room for noise our text as words like *photograph* are accepted.

If those words that are not terms are compare with those word that are terms we realise that the difference is to be found in the lexical base to which they are attached. Hence it can be said that these general-sense forms only turn into terms if the lexical base is a term. Therefore to improve the extraction module it is necessary to test in an ontology whether the lexical base has a specialised meaning or not.

Finally, this tool still allows for false positives due to the homonymy of some forms like *metr-* meaning 'measurement', 'device', 'measurement procedure' or 'ureter'. This ambiguity can be tackled by analysing the constraint posed in the combination of these homonymic forms.

7. Conclusions and future development

This paper has attempted to show the uselfuness of neo-classical forms in term extraction within the biomedical domain. It has been done through the implementation, in a more complex system, of a term detection module containing any Greek and Latin form. Therefore we have tried to show in which way morphological features of terms are useful for their automatic extraction in a particular specialised domain.

As for the module of neo-classical form detection, it is noteworthy that it is based on the very same strategy used by specialists when term composing and term decomposing within its specialty. Besides, it is noticeable that this module can be combined with other complementary ones.

The good results of this tool as far as recall and precision is concerned given an account of its uselfuness and permit to foresee a future fields of work:

- establishment of constraints regarding form combination
- analysis in further detail term detection with neo-classical suffixes and prefixes of a general sense, establishing a hierarchy of its pertinence in medicine.
- further study of the combination of a form with general-language word, combining the module of neo-classical form detection with an ontology.
- test these results in larger corpora.
- integrate the module of neo-classical form detection in the extraction of polilexical terms.

8. Aknowledgement

This paper has been written within the research project *Scientific and technical terminology: recognition, analysis and retrieval of formal and semantic information* (PB-96-0293), supported financially by the Spanish government. We would also like to thank J. Morel for his unvaluable help in translating this paper from Catalan into English.

9. References

Bernabeu, J. Et al. (1995) El llenguatge de les ciències de la salut: introducció ala formació de termes mèdics. València: Generalitat Valenciana.

Bourigault, D. (1994) «LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes». PhD

[3] This compound allows terms like *embolism*.

Thesis. Paris: École des Hautes Études en Sciences Sociales.

Cabré, M.T. (1999) La terminologia. Representación y comunicación. Una teoría de base comunicativa. Barcelona: IULA, Universitat Pompeu Fabra. (Sèrie Monografies, 3).

Cabré, M. T., Estopà, R. and Vivaldi, J. «Automatic term detection: a review of current systems». To appear in D. Bourigault, C. Jacquemin, M.-C. L'Homme (eds.) Recent Advances in Computational Terminology John Benjamins.

Estopà, R. (1999). Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'extracció automàtica de candidats a unitats de significació especialitzada). Barcelona: Universitat Pompeu Fabra. These doctoral.

Estopà, R., Vivaldi, J. and Cabré, M. T. «Extraction of monolexical terminological units: requirement analysis». Paper submitted to: Computational Terminology for Medical and Biological Applications, Patras, Greece. http://www.iula.upf.es/iulaterm

Frantzi, K. T. (1997). «Incorporating context information for extraction of terms». Proceedings of the ACL/EACL, Madrid: 501-503.

Kageura, K.; Umino, B. (1996). "Methods of Automatic Term Recognition". Terminology, 3:2. 259-290.

López Piñero, J. M. and M. L. Terrada Ferrandis (1990). Introducción a la terminología médica. Barcelona: Salvat Editores.

Maynard, D. and S. Ananiadou (1999). «Identifying contextual information for multi-word term extraction» In Sandrini, P. (ed.) TKE '99: Terminology and Knowledge Engineering. Vienna: TermNet: 212-221.

REALITER (1997). Taula de formants cultes. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Smith, G. Et al. (1996) Quick medical terminology. Pittburgh: John Wiley & Sons.

Vivaldi, J. and Rodríguez H. (2000). "Improving term extraction by combining different techniques". Paper submitted to: Computational Terminology for Medical and Biological Applications, Patras, Greece. http://www.iula.upf.es/iulaterm.

Vossen, P. (1998). EuroWordNet: a multilingual database with lexical semantic networks. Computers and the humanities. Flushing, NY: Queens College of the City University of New York, 1966-. Vol. 32, no. 2-3.