Extraction of monolexical terminological units: requirement analysis

Rosa Estopà, Jordi Vivaldi, Teresa Cabré
Institute for Applied Linguistics, Universitat Pompeu Fabra
Rambla Santa Mónica, 32
Barcelona, Spain, 08002
rosa.estopa@trad.upf.es jorge.vivaldi@info.upf.es teresa.cabre@trad.upf.es

Abstract

Specialised texts contain both polilexical and monolexical terminological units. Monolexical terms are not treated in most current extraction systems mainly due to their high degree of polysemy. However this is mainly true, in a specialised domain such as medicine it needs further explanation. In this paper we discuss the requirements posed by term extractors to detect monolexical terms. In addition, we present some results obtained by applying our requirements to a medical corpus.

1 Introduction

In specialised texts polilexical terminological units (PTU) appear together with monolexical ones. However, current extraction systems do not usually attempt to recognise monolexical terminological units (MTU) and when they do so is because a MTU is part of a polilexical unit. Actually most terminology extraction systems are exclusively devoted to PTU detection (Cabré et al.).

It has been argued that this restriction is due to the fact that formally MTU do not contain specific structures and MTU's degree of polysemy is higher than that of polilexical ones (Daille, 1994; Jacquemin, 1996; Naulleau, 1998, etc). However this is mainly true, in a specialised domain such as medicine it needs further explanation.. Hence, provided that MTU are more idiosyncratic and polysemous than PTU, it can be shown that monolexical terms from the medical domain have morphological, semantic and pragmatic properties that are related to their corresponding conceptual class and that can be useful for automatic extraction.

Taking into account morphological, semantic and contextual features of MTU as a starting point this paper aims at discussing the requirements needed by the extraction techniques with regard to MTU. Also it discusses the results from applying three techniques to a medical corpus.

2 Extraction techniques and monolexical terminological units

Following (Cabré's 1999) definition of terminological unit we consider that a MTU is any lexical unit found between blanks that is used in a specialised meaning within a given text. However, not all the MTU have the same morphological structure. In medical texts three main types of monolexical terms can be distinguished: simple (*heart*, *nodule*), inflected (*inflammation*, *tumefaction*) and, most importantly, neo-classical compounds (*hemorrhage*, *pnemonitis*). According to these different structures, we start with some hypotheses so as to attain the most adequate and efficient MTU extraction strategies from biomedical texts:

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

b) Taking into account that there are linguistic differences, we should avoided using a single technique for all MTU. Consequently, we state that technique combination is the most useful strategy to follow in order to benefit from them.

2.1 Monolexical terms comprising neo-classical forms

It is well known that medical vocabulary is based on Greek and Latin forms (stems, prefixes and affixes). As many scholars have pointed out (Quintana, 1989; López Piñero and Terrada Ferrandis, 1990; Love and Davis, 1990), around a thousand of Greek and Latin stems and 80 related affixes account for most of the medical vocabulary. Using this data together with a set of combination rules a particular module could 'reproduce' specialists strategies to recognise this type of units. Thus some approaches within the terminology extraction field have been carried out, for instance (Spyns and De Moor, 1998), (Ananiadou 1994) and (Zweigembaum and Grabar, 1999).

A module from the extractor could extract neo-classical compounds in a way similar to that of specialists in the medicine. To do so it could make use of a dictionary of 1100 Greek and Latin forms from the related domain together with a small number of form combinations rules. Such module yields to detect a great deal of MTU as well as UTP since monolexical TU are often hyperonymous of polilexical TU.

Together with neo-classical compounds, in medical text there can be found units from anatomy, zoology, botanics, bacteriology and virology. Nomenclatures from these areas, based on well-established classifications, are reached by international agreement and they all have fixed endings (for instance, in biochemistry the *-in/-ine*, *-ol/-ole* and *-ase* suffixes are highly productive: *aldolase*, *cloramfenicol*, *tetracycline*, *prostanglandin*, etc.). Thus, however the structure of such units is not as well-defined as that of neo-classical compounds, the treatment of the above-mentioned suffixes could help to achieve better results in term extraction without using a terminological dictionary.

2.2 Simple monolexical terms

Biomedical corpus analysis shows that the number of simple terminological units (STU) is lower than that of other specialised units (Estopà, 1999). However, it is interesting to extract from texts all STU because they are referred to medical basic concepts mainly referred to anatomy and pathology. Often they are the lexical base of inflected meaning units:

blood → *bloody, bloodless*
bone → *bonelet*
nerve → *nerval, nervous, nervosity*
grippe → *grippal*, *postgrippal*
nodule → *nodulation, nodular, nodulated*
ulcer → *ulcerate, ulcerous, ulcerative*

Further they are head of many syntagmatic units as:

blood → *cord blood, lacky blood, occult blood, blood plasma,* etc
bone → *flat bone*, *lacrimal bone*, *palatine bone*, etc.
nerve → *anabolic nerve, exodic nerve, vasomotor nerve, Wrisberg's nerve,* etc.
grippe → *Balkam grippe, devil's grippe, Dabney's grippe,* etc.
nodule → *Kerkring's nodule, pulp nodule, triticeous nodule, nodule of vermis,* etc.
ulcer → *Aden ulcer, atonic ulcer, chronic ulcer, gastric ulcer, Hunrer's ulcer*, etc.

STU are difficult to be automatically recognised since their specialised nature is absolutely particular. Thus STU do not show explicit morphological and syntactic features contributing to their automatic detection. Hence access to STU can only be gained by using lexicographic and/or contextual strategies.

Moreover, STU automatic detection is even more difficult if we take into account that they can be given more than one semantic value. In other words, STU are often highly polysemous units and, devoid of context, it is terribly difficult to know whether their meaning is specialised or general (Cabré, 1999). Here only context could help to disambiguate the specialised nature of STU although it has been noted that a high percentage of polysemous terminological units refer to parts of the human body and related pathological signs.

Therefore it can be concluded that two combined strategies help in STU detection: the use of an ontology and/or the analysis of STU linguistic context.

2.3 Inflected monolexical terms

MTU analysis from medical contexts allows to state that, from a lexicomorphological viewpoint, these units are formed from a lexical stem always being a specialised meaning unit that is related to either a simple term or a neo-classical form relevant in medicine. Also it can be said that these stems are combined with affixes whose number is much lower than that of affixes combined with non-specialised lexical stems. For instance *-tion* is highly productive within the semantic class of actions and operations: *infection*, *inoculation*, *vaccination*, etc.

According to the above considerations, it seems reasonable to say that an efficient MTU extractor should relate all TU sharing the same lexical stem and that these TU will be related to either a simple TU or a Greek and Latin form.

A more lexicographical-based strategy to detect inflected MTU would consist, similarly as in STU, in looking for them in an ontology containing semantic classes relevant in medicine. This strategy can also be combined with context analysis.

In conclusion we believe that linguistic features particular to different types of MTU should lead to techniques allowing their extraction.

3 A tool for detecting monolexical term candidates

According to the previous statements, it seems clear that each type of unit requires a specific analysis. As it will be discussed, it is not practical to propose a single strategy covering all kinds of MTU. Rather, it is better to use a set of modules analysing each candidate from different perspectives and, according, a term extractor is being developed at out institute. In this section we will provide a short description of the modules used in this tool. Further detailed information can be found in (Vivaldi and Rodríguez, 2000).

Detection of Greek and Latin compounds (GLC). This module requires term candidate decomposing so as to obtain their components' meaning (López Piñero and Terrada Ferrandis, 1990). However simple this task may be some prevention about false positives should be taken. False positives refer as to general-language words that can be decomposed in Greek and Latin forms but the outcome has no sense whatsoever. Experiments using this module have shown a high precision but a relatively low recall. Besides, this technique does not require investing expensive resources to lexicon enlargement. See (Estopà et al., 2000) for more details.

Semantic data extractor. To benefit from an ontology it is first necessary to gain access to a specialised resource that can provide a semantic tag for each term candidate. Thus we have used the European WordNet Database (EWN DB)[1] to determine whether words belong to the medical domain or not. Although this resource mainly includes general-language vocabulary it has a strikingly great coverage of the domain of medicine. Besides, we have extended its coverage to nouns, adjectives and, even, to the adjective-noun semantic relation, which has been defined but still not used. In addition, it has to be considered that specialised medical thesaurus are not easily found in languages other than English. We consider that it provides an adequate support to our technique.

Context analyser. Our module's approach is based on (Franzi, 1997) and (Maynard and Ananiadou, 1999), subject to the resources available at our lab and adapted with additional experiments. The bases for context analysis are: (i) words surrounding "prime" terms[2] candidates can become useful signs for other terms and (ii) among such context words, "prime" terms that are semantically similar to the considered term candidate provide additional information.

The whole process comprised processing steps that are common to most NLP systems: segmenters, text handling, morphological analysis and POS tagging. Then a candidate extraction module, based in (Bourigault, 1994), seeks all possible terminological sequences that will be analysed by the above-mentioned term analysers.

4 Experiment test

The above three techniques have been applied to a highly specialised medical corpus on Rickettsial diseases from Farreras-Rozman's *Medicina interna* (1997). A term candidate extractor, based in (Bourigault, 1994), proposes 645 units as MTU candidates of 12.069 occurrences. To evaluate the performance of automatic techniques we have taken into account the manual terminological retrieval made by three specialists in the domain.

5 Result analysis

From all the proposed MTU, 339 (51,83%) have been selected by specialists. The results of this selection and the application of the three extraction methods on the same corpus are the following:

[1] 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).

[2] A "prime" term is candidate that has been found in the EWN as having one or more medical meanings.

MTU		Detection Method							
		Manual		EWN		GLC		Context	
		%	#	%	#	%	#	%	#
Neo-classical compounds		33.6	114	12.9	40	29.5	100	22	74
Standardised Nomenclatures	Biochemistry	9.7	33	1.2	4	-	-	3.2	11
	Zoology	7.4	25	-	-	-	-	4.7	16
Simple		28.3	96	19.8	67	-	-	20	68
Inflected		20	68	1.8	9	-	-	8.3	28
Inherited compound		0.9	3	-	-	-	-	0.6	2
Total		100	339	34.8	120	29.5	100	58.8	199

Table 1. Term candidates obtained by each method

As it can be seen in the above table[3] the first column shows all kind of MTU selected by specialists. The second column indicates the number of MTU that have been selected by specialists according to unit types. The third column accounts for the number of MTU that have been detected using the ontology method and, finally, in the fourth column the context strategy is represented.

As table 1 indicates, results from the application of the three extraction methods show what follows:
1. However the EWN module is the one that detects less units, data analysis shows that this is a very reliable module as it hardly produces noise[4].
2. The module of neo-classical form extraction is very useful to detect Greek and Latin compounds as it detects 88% of these units found in a text. Besides, it hardly produces noise.
3. Finally, however the context module is the one that leads to detect more number and various types of MTU it is the one that produces more noise[5].

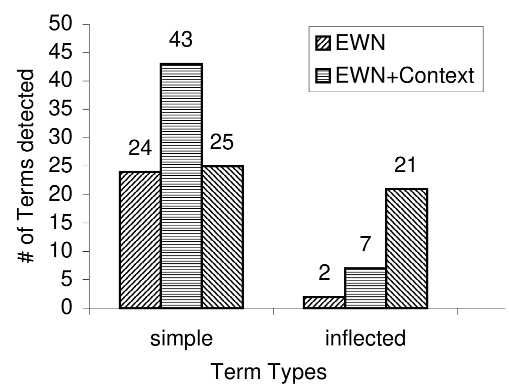
To test our starting standpoints we should consider the degree of overlapping in the results from detection among the above three methods. As can be seen from the above tables, there is a low degree of overlapping among the three methods, which means that the three techniques detect different MTU. Thus it can be inferred that an extractor needs to use these three methods combinately to lead to a higher recall figure.

Taking into account the results obtained with all three methods, in order to test our hypothesis it is important to analyse the overlapping degree existing in the detection stage. This can be observed in Table 2 and Table 3 below. Table 2 and Graph 1 show the results obtained for simple and inflected MTU as well as the standardised nomenclatures. Table 3 and Graph 2 only account for MTU formed by Greek and Latin forms.

MTU type	Detection Method			Total MTU detected	Over lapping
	EWN	EWN + Context	Context		
Simple	24	43	25	92	43%
Inflected	2	7	21	30	23%

Table 2. MTU detected (part 1)

Graph 1. Overlapping in the detection of simple and



inflected MTU and nomenclatures

Table 2 and Graph 2 show that our proposal is largely supported regarding simple terms: more than half of such terms are detected only by one method: 24 terms

---

[3] By inherited compounds, we mean those compounds formed by lexical units from the language. For intance, *escalofrío* ('shiver') in Spanish and *abaixallengües* ('spatula') in Catalan.

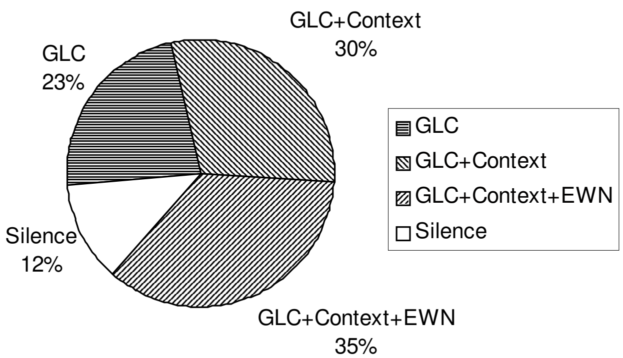
[4] Results obtained by this method are the following: precision: 94%, recall: 27% for all monolexical nouns with a single meaning.

[5] This method provides a list of all candidates ordered according to its context factor. In the first 9% of the whole candidates with all its experienced variants, the best precision and recall results are the following: 79% and 23% respectively.

are detected by the EWN method and 25 terms are detected by the context method. It does not apply to inflected terms due to the following two reasons:
Inflected MTU are usually derived from verbs. This kind of relations has been defined yet not applied. So it may happen that both the verb and its nominal derivative are found in the same database there is no relation whatsoever between them[6].
The helpfulness of the context method is subject to the features of the specific occurrences as well as the number of occurrences.

	Detection Method		
	GLC	GLC+Context	GLC + Context + EWN
#	26	34	40

Table 3. MTU detected (part 2)



Graph 2. Overlapping in the detection of GLC concerning MTU

As can be seen from Table 3 and Graph 2 the method used to detect GLC is pretty efficient (88% of recall) but it should be noted again that the percentage of terms detected by all the three methods is low (35%). This is so because the module responsible for GLC detection is very specific.
Let us see some cases of MTU detected through the proposed strategies:

(1) *hepatitis* — *Otras formas crónicas pueden ser osteomielitis,* ***hepatitis****, infección de aneurismas aórticos y, en ocasiones, de material protésico.*

(2) *vesícula* — *A los 26 días d el inicio de la fiebre, aparece el exantema, maculoso y luego papular, en cuyo centro se desarrolla una* ***vesícula***

(3) *acufeno* — *El 25,30 % de los pacientes refieren tos seca, mareos, acufenos, fotofobia, náuseas, dolor abdominal y estreñimiento.*

(4) *ingle* — *Suele localizar se en zonas cubiertas y de flexión (axila,* ***ingle****, zona interdigital, pubis, región retroauricular, cuero cabelludo, zona submamaria, hueco poplíteo), lo que dificulta su detección.*

[6] EWN already foresee this relation but it is not filled.

(1) is the most positive case, although it is not a frequent term in the analysed text (2 occurrences), it is supported by all three methods: *a)* as for ontology it is assigned the semantic label *disease; b)* it can be detected through the neo-classical form extractor: *hepat-* ('liver') and *–itis* ('inflammation of a body part'); *c)* within the text it is surrounded by other diseases or pathological signs (*osteomilitis*, *infección de anuerismas aórticos*).

(2), (3) and (4) account for simple MTU that, because of their particular linguistic and contextual features, are supported by different methods. In (2) *vesícula* ('vesicle') appears only once throughout the corpus and its context is not highly specialised. As a result, it is only supported by EWN wherein the semantic label *body part* is encountered. In (3) acufeno ('*acousma* ') is not found in the ontology and it only appears once. However it has been detected because its context is terminologically very rich: *tos seca*, *mareos*, *fotofobia*, *náuseas*. Finally (4) shows overlapping of methods since *both the context and the lexicographical source support ingle ('inguen').*

It is noteworthy that of the whole simple terms found in the text (96) less than half of them (43) are recognised by the two available methods (EWN and Context). It means that only one of these two methods detects the rest and there are 4 remaining MTU that are not detected because they are not encountered in the ontology and their context is terminologically poor.

6 Conclusion

In this paper we have discussed the efficacy of using more than one strategy regarding automatic extraction of MTU through an experiment test in the medical domain. We have shown that the results obtained lead to the following considerations:

1. Linguistic features of TU give rise to different extraction techniques.
2. The combination of more than one differential method is relevant so as to attain a higher degree of recall.
3. Each specialised domain can choose the most suitable strategy. Thus in medicine we have shown how efficient is a method based on neo-classical forms. This same method would not be efficient in a

technical field such as computer science or automotion.

Our further research will be devoted to the topics listed below:
- Improving the system by refining the three methods, especially that of context techniques.
- Relating deverbal nouns to their corresponding verbs in order to reach better precision in the detection of inflected MTU such as *infección* ('infection'), *intervención* ('intervention') and the like.
- Test our hypotheses with other patterns in larger texts.

Acknowledgements

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 like to thank J. Morel for his unvaluable help in translating this paper from Catalan into English.

References

Ananiadou S. (1994). «A methodology for automatic term recognition». *Proceedings of the Coling 94*, Kyoto, Japan: 1034-1038.

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

Daille, B. (1994) *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Université Paris VII: PhD Thesis.

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

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)*. PhD thesis, Barcelona: Universitat Pompeu Fabra.

Estopà, R.; Vivaldi, J.; Cabré, M. T. (2000). Use of Greek and Latin forms for term detection. To appear in Proceedings of LREC 2000. May 29 - June 1. Athens.

Jacquemin, C. (1996) "What is the tree we see through the window: a linguistic approcah to windowing and term variation". Information Processing and Management, 32/4, 445-458.

Quintana, J.M. (1989). La terminología médica a partir de sus raíces griegas. Madrid: Dykinson.

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

Love, G. and P. Davis (1990). Curso rápido de terminología médica. México: Limusa.

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.

Naulleau, E. (1998) *Apprentissage et filtrage syntatico-sémantique des syntagmes nominaux pour la recherche documentaire*. Université Paris VIII: PhD Thesis.

Spyns,P and G. De Moor. (1998) Robust recognition of unknown Dutch medical vocabulary. Division of Medical Informatics, State University Gent. Belgium. (http://allserv.rug.ac.be/~pspyns)

Vivaldi J. and H. Rodríguez (2000) «Improving term extraction by combining different techniques». Proceeding of the NLP2000 workshop on "Computational terminology for medical and biological Applications". Patras, june, 4th.

Vossen, P. (1999) *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.

Zweigenbaum P. and N. Grabar (1999). Automatic acquisition of morphological knowledge for medical language processing. In Werner Horn, Yuval Shahar, Greger Lindberg, Steen Andreassen, and Jeremy Wyatt, editors, Artificial Intelligence in Medicine, Lecture Notes in Artificial Intelligence, pages 416-420. Springer-Verlag.