A simple and robust method for extracting terminology

LUIS SARMENTO
University of Porto, Porto, Portugal
las@letras.up.pt

RÉSUMÉ
Nous présentons une méthode simple, mais efficace, pour l'extraction de terminologie à partir de textes techniques. La méthode est basée sur l'observation que dans les domaines techniques il est beaucoup plus simple de déterminer ce que n'est pas une unité terminologique valide que d'identifier ce qui est probablement une unité terminologique. Notre méthode se fonde sur un ensemble de filtres qui excluent les pluritermes selon des règles simples concernant leur contexte et leur structure lexicologique interne, et elle n'exige aucun prétraitement spécial comme le POS *tagging*. Des règles sont été codées manuellement par un simple procès incrémental et elles peuvent être écrites en plusieurs langues sans effort. En plus, la méthode peut traiter plus de deux millions de mots par minute avec un ordinateur standard. Bien que la méthode ait été originalement prévue pour l'extraction terminologique semi-automatique, nous croyons qu'elle peut également être appliquée à une procédure complètement automatisée, la rendant appropriée à l'extraction d'information à grande échelle. Nous commencerons par expliquer notre motivation principale pour établir cette méthode et nous décrirons son rôle dans sa plus grande portée, le *Corpógrafo*. Nous présenterons la façon d'établir la méthode courante, dès les premières approches de la version en cours, précisant les problèmes rencontrés à chaque étape. Nous donnerons alors les résultats de la version actuelle de la méthode d'extraction en corpus de domaine spécifiques en anglais. Enfin, nous présenterons des propositions pour l'avenir et expliquerons un petit lexique sémantique facilitant les procédures d'extraction d'information à grande échelle.

ABSTRACT
In this paper we will present a simple, yet effective, method for extracting terminology from technical text. The method is based on the observation that for technical domains it is much simpler to describe what a valid terminological unit cannot be than what it can possibly be. Our method relies on a set of filters that exclude multi-word units according to simple rules regarding their context and internal lexical structure, and it does not require any special pre-processing such as POS tagging. Rules were hand-coded in a simple incremental process and may be ported to several languages with little effort. Additionally, the method is able to process more than two million words per minute on a standard computer. Although the method was originally intended for semi-automatic terminological extraction, we believe that it can also be applied in fully automated procedures, making it appropriate for large-scale information extraction. We will start by explaining our main motivation for building this method and we will describe its role in a larger framework, the *Corpógrafo*. We will then present the process of building the current method, from the first very simple approaches to the current version, pointing out the problems encountered at each step. We will then present results of applying the current version of the extraction method to specific domain corpora in English. Finally, we will present future plans and explain how we are currently in the process of building a small semantic lexicon for helping future large-scale information extraction procedures.

MOTS-CLÉ/KEYWORDS
terminology extraction, method, technical writing, translation, information

Introduction

In recent years there has been an exponential growth of multilingual written information, which has revolutionized the way we process information. The importance of terminology has grown in a similar fashion due to the ever-growing need for terminological resources. However, terminology is now facing several new and difficult challenges, which are changing the traditional way of thinking about terminology and terminological resources. Terminology, as a science, is facing new requirements and needs, especially those coming from the fields of information retrieval, information extraction, and translation. In fact, terminology has widened its scope moving from large centralized efforts to much more dynamic scenarios related to technical writing, translation, information indexing or retrieval. If we consider, for example, the terminology needs of a freelance translator or of any small-scale translation company that cannot specialize in a specific domain of knowledge nor afford the luxury of consulting renowned domain specialists, we see that the place for terminology is essentially that of helping to produce dynamic translation dictionaries for everyday activities.

In these cases, the source for the required terminology is mainly the incredibly large amount of text that is available on line in many languages, whose reliability is not always possible to validate. For such a fast resource production cycle, the appropriate terminology extraction methods need to deal robustly with large volumes of text and, at the same time, with many different knowledge domains, and possibly even with different languages.

Practical Terminology Extraction

The need to operate on different knowledge domains brings further difficulties. The first one is that extraction algorithms can no longer easily rely on specific-domain heuristics, such as internal evidence (e.g.: suffixes or prefixes) or the knowledge of basic terminology of the field, to identify other valid terminological units (TUs). Extraction algorithms need to be able to identify "RISC processor" as much as they identify "1,3 BUTADIENE". In fact, in an open domain, virtually anything can be a valid TU, which requires algorithms to identify as valid terminology even items that are usually considered general language. To complicate matters further, general language has been adopting many technical terminology units, which are now part of our every day language, so the difference between technical terminology and general language is not always clear.

The definition of a valid TU is thus extremely difficult to formulate. We may say TUs are lexical representations of relevant concepts of a given field, but this definition does not help us enough when we are dealing with automatic terminology extraction. A more practical formulation, which is conceptually simplistic but certainly more helpful from the point of view of automatic terminology extraction, is that valid TUs may be found among the noun phrases of a technical text. Such noun phrases may represent objects, processes, events, and other relevant concepts of a certain domain. The noun phrases collected in a technical text may, however, include many other very frequent constructions, such as certain collocations and some typical or stylistic expressions that cannot be considered TUs (e.g.: "scope of the work" or "main problem"). Named-entities may also be considered valid TUs, such as in the case of the names of system/tools ("Windows XP"), or the inclusion of proper names in TUs (e.g.: "Fibonacci number"), but names of people, places or organizations are not usually considered terminology. Verb phrases are usually not considered TUs, as they can be derived from the corresponding noun phrases.

One can also expand this notion of terminology to a broader scenario and consider other cases. If we apply the previous definition that states that terminology is related to the representation of concepts of a given field to journalistic text, then we are forced to consider all named-entities as valid TUs, which approximates terminology extraction of named-entity recognition (Sekine 2004). This may look odd, but in fact named-entities may be as problematic to translators as "standard" technical terminology (consider for example the proper names of certain administrative political or governmental organizations).

As we have explained before, from the practical point of view of translation or technical writing, the web is now one of the major sources of information for terminology extraction. By

using a small set of known seed TUs and a standard web engine, one may now easily create a technical corpus, consisting of several million words. Systems that help users building huge corpora automatically starting from the same seeds are also available (Baroni and Bernardini 2004), allowing users to collect a vast amount of information for subsequent terminology extraction, which offers both challenges and opportunities. On the one hand, extraction methods need to be fast enough to process megabytes of text in a few minutes in order to cope with rapid development cycles. In practice, this may leave out any method that requires long pre-processing times. On the other hand, because the web is a very redundant source of information, it is possible to focus on alternative data-driven methods, which may achieve good results with reasonably simple and fast algorithms. Such data-driven methods are being successfully used in several semantic-based tasks, such as concept discovering or named-entity recognition and classification (Pasca 2005). Data-driven methods are also more robust in coping with real-life text sources that nearly always have very complex formatting, which greatly complicates the plain text extraction required for further deep parsing processing (e.g.: extraction problem in complex HTML or PDF files).

Simultaneously, there is an increasing interest in the development of algorithms capable of operating reasonably well in several languages. Aligned terminological resources are particularly valuable for translation (both manual and automatic) and cross-lingual information extraction. However, developing a terminology extraction system for one language is already difficult, time-consuming and may involve producing a set of language-dependent extraction rules, which are usually difficult to port to other languages.

There are some methods that conform to all, or most, of the previously mentioned requirements. Two of such methods are described in (Merkel and Andersson 2000) and follow what the authors call a "knowledge-lite" approach to terminology extraction. Both of the described methods operate on untagged text and start by collecting lists of candidate multi-word TUs, either n-grams with a minimum pre-defined size or word sequences with significant lexical cohesion based on entropy filtering. Incorrect candidates are then filtered out using one of two possible sets of rules that impose restrictions on the words that may start, end or exist inside a valid TU. Candidate TUs are simply checked against lists of prohibited words that the authors compiled manually. In a final step, a frequency threshold is applied to remove the least frequent candidates, which helps to reduce the noise. The authors report very high precision levels in detecting terminology from technical manuals, especially using one of the two possible sets of filters. Following a similar approach, we have developed a method that extends the previously described one and is able to extract TUs of any length, even single-worded ones, with high precision-levels.

Motivation

Our interest in terminology extraction is related with the development of *Corpógrafo* (Sarmento et al. 2004; Maia, forthcoming), an online environment for performing linguistic studies and developing terminological resources. *Corpógrafo* enables users, mainly lexicographers and translators, to build their own corpora and to extract terminology, definitions and semantic relations. *Corpógrafo*'s main goal is to accelerate as much as possible the work that users would do manually. *Corpógrafo*'s production cycle is centered on semi-automatic approaches which require some user intervention to ensure the quality of the resulting resources. As far as terminology extraction is concerned, *Corpógrafo*'s role is to suggest good candidates for valid TUs to users, taking into account that validation should be simple and fast. When a TU is validated, it may be stored in a dedicated database and related information can then be extracted, if desired, namely definitions and possible semantic relations.

Corpógrafo has very special needs regarding text-processing techniques, especially terminology extraction methods. First of all, *Corpógrafo* is intended for multilingual work centered around Portuguese, therefore it needs to be able to extract terminology in other languages in which our community of users is interested, namely English, Spanish, French, Italian and German (and hopefully even more). Therefore, terminology extraction methods need to be easily implemented and ported to many languages. Also, *Corpógrafo* is not restricted to any particular knowledge domain and needs to have broad scope terminology extraction methods. Users have used

Corpógrafo to develop terminological databases in various domains, from Natural Language Processing (Oliveira et al. 2005) to Neurology.

All these conditions impose very rigid restrictions on the nature of the possible terminology extraction methods. Since we were taking into account the need for validation, we started making some simple experiments to verify how *Corpógrafo* users responded to very simple extraction methods. We implemented a simple extractor that produced lists of n-grams ordered by their frequency in the corpus. Users chose the size of the n-gram and *Corpógrafo* produced the corresponding ordered list. Surprisingly, users considered such a simple method to be of help for building glossaries. The fact was that for n-grams of length 3 or more, the most frequent n-grams were in fact valid TUs because of the very frequent Noun + preposition + Noun form in Portuguese. For n-grams of size 2, the users also considered the results reasonably efficient. Similar results were obtained for other languages, and despite their inherent limitation, this simple extraction method was considered fairly appropriate for a semi-automatic environment.

Improving the simple approach

The first improvement consisted in constraining the possible POS structure of the candidate n-grams. Traditionally, only noun-phrases are usually considered valid TUs. There are multiple possibilities for terminologically valid noun-phrases, but a simple way to verify that we are probably dealing with a valid structure, for each extracted n-gram, is to check if the starting and ending words of the n-gram are nouns or adjectives. We implemented such a mechanism by using a simple dictionary lookup. All n-gram sequences that conformed to these conditions would be considered valid, while the others would be discarded. We were able to obtain significant improvements in precision in relation to the previous basic approach, but there was a significant problem. Any valid TU that started or ended by a word not present in the dictionary would never be considered. This is especially severe in some domains where there is a great probability of proper names occurring, such as in "Alzheimer's disease" or acronyms, and that almost surely would not be included in general language dictionaries like the one we had available. Additionally, we were only able to find a high coverage dictionary for Portuguese and so porting this method for other languages would be impossible with our current resources.

The solution for this problem came from a simple observation: apparently it is much easier to exclude invalid TUs than it is to select valid ones. This goes exactly in the opposite direction of most extraction algorithms that try to identify TUs based on linguistic or statistical descriptions of what they should be. Such descriptions, however, are not stable among domains of knowledge, whereas it is possible to formulate simple lexical-based rules about what a valid terminological is not. For example, it is possible to say with a great degree of probability that a valid multiword TU will not start with "with". We are not arguing that there is no valid TU at all that starts by "with", because there might be a particular domain where "with" is, in fact, a valid TU. What we are saying instead is that such a simple lexical-based rule is verifiable in the great majority of domains and it may, therefore, be used to exclude invalid candidates with precision. The same may be said for certain words that may never end, or even be part of any TU. All these rules require is simple string matching against elements belonging to three lists of single words:

1. **List of Non-Starters**: words that may never start a valid TU. E.g.: "a", "about", "across", "after", "along", "an", "at", "away", "be", "between", "by", "for", etc.

2. **List of Non-Enders**: words that may never end a valid TU. E.g.: "a", "about", "after", "along", "an", "at", "be", "become", "becomes", "developing", etc.

3. **List of Prohibited words**: tokens that may never occur in a valid TU. E.g.: "above," "according", "across", "actual", "actually", punctuations, symbols, etc.

We, thus, implemented in Corpógrafo, a very simple extraction algorithm:

1. Let the user define S, which will be the size in words of TUs to be found. L is the currently empty list of candidates.

2. Transverse the entire text, on a word-by-word basis. For each word:
   a. Build an n-gram N of length S starting by the current word and composed by following S-1 words;
   b. Check n-gram N against the 3 lists - Non-Starters, Non-Enders and Prohibited words;
   c. If the n-gram is not excluded then store or update n-gram count in the list L.

3. Order list L by frequency and present results to the user.

The exclusion lists were manually built in an iterative fashion. We started with empty lists and ran the algorithm over a test corpus in order to check the resulting n-gram list. We would then add words that could be safely included in any of the lists to exclude some of the wrong candidates. The extraction procedure was then run again to check if the filtering had achieved the intended effect and also to choose more words to be included in the lists. The process was repeated several times until results were found satisfactory. We used a corpus of texts about neurology containing 29192 tokens to train the system, i.e. to obtain the exclusion lists.

For the following tests, we used a control corpus containing 109 documents written in English about computational processing of Portuguese, for a total of 498226 tokens. For now, we will only be considering n-grams of size 2. Results produced for n-grams of size one are obviously noisier and we will be presenting a more appropriate method in the next section. Results regarding n-grams of size 3 are naturally better in precision as noise becomes less significant, so we will be focusing on n-grams of size 2. The system generated 51012 candidates of size 2. Table 1 contains the top 30 most frequent size 2 n-grams, along with the corresponding number of occurrences in the corpus:

Candidates (1-15)		**#**	**Candidate (16-30)**		**#**
1	binding constraints	370	16	ele próprio	70
2	natural language	218	17	anaphoric expressions	68
3	et al	215	18	Computational linguistics	68
4	binding theory	152	19	anaphoric links	63
5	times new	120	20	long-distance reflexives	61
6	new roman	120	21	constraint grammar	60
7	anaphor resolution	119	22	test set	59
8	brazilian Portuguese	107	23	definite descriptions	58
9	speech recognition	102	24	Reference marker	57
10	semantic representation	98	25	information retrieval	57
11	reference markers	89	26	obliqueness hierarchy	57
12	reference processing	88	27	Semantic prosody	57
13	machine translation	85	28	binding principles	54
14	european portuguese	82	29	language identification	54
15	language processing	74	30	word class	54

Table 1 - The top 30 most frequent size 2 n-grams, obtained by processing a 498226 tokens corpus using the improved method.

After manual validation, each candidate was classified according to three possible categories:

A – Correct and Valid. All candidates that are a valid TU;

B – Possibly Correct. Any candidate which can be considered a valid TU under certain circumstances, but that could be excluded using more rigid criteria (names of people, incomplete but possible, etc.);

C – Incorrect or malformed. Absolutely incorrect.

We then calculated the precision of the process considering the top 30, top 50 and top 100 most frequent n-grams:

	Class A	Class B	Class C	P (A only)	P (A + B)
Top 30	22	4	4	70.0%	86.7 %
Top 50	36	5	9	72.0 %	82.0 %
Top 100	75	10	15	75.0 %	85.0 %

Table 2 – Values obtained regarding the precision of the improved extraction method.

As we can see, results are reasonably good as the precision of the process became high enough to help users collect more TUs in less time. Precision is at least 70% and, in some cases, it might reach more than 80%, if we relax some of the requirements for a candidate to be a valid TU. Results were, however, inferior to those reported in (Merkel and Andersson 2000) who claim their method can obtain precision levels of more than 90%. Nevertheless, we were able to obtain very similar results in all of the target languages we were interested in: Portuguese, English, French, Spanish and Italian (we were not able to obtain good results for German).

Improving the results with additional context filtering

After analyzing the results more thoroughly, we noticed that incorrect (Class C) and possibly incorrect (Class B) candidates were mainly obtained under the following circumstances:

1. The candidate is part of some very frequent multiword units, such as a frequent collocation or a phrasal verb that was not possible to exclude using the filters. E.g. “carried out”. Adding new words to the exclusion lists could have a negative impact on the extraction’s recall by excluding many other valid terminological entries, and would simply make the lists too long to be manageable.

2. The candidate is an incomplete valid TU. For example, for n-grams of size 2 we found many situations where the candidate obtained is, in fact, the first or the last 2 words of a valid 3 word TU. E.g.: “natural language” and “language processing” instead of “natural language processing”.

3. The candidate is part of a frequent and genre dependent convention (ex: citations). E.g.: “et al”.

4. The candidate results from problems in pre-processing at the text extraction or at the tokenization stages. E.g.: “times new”, “0cm-0cm”.

While the second problem seems to require a more complex solution, imposing an additional restriction could solve problems 1, 3 and partially 4. If we consider that relevant TUs are mostly noun-phrases, we may increase precision without sacrificing the recall too much, by only considering candidates that are preceded by certain particles, such as articles, quantifiers, prepositions, some verbs and some punctuation. We thus defined another list including words that must precede an n-gram for it to be considered a valid TU, such as: “a”, “all”, “an”, “and”, “as”, “another”, “both”, “by”, “called”, “each”, “few”, “for”, “like”, “many”, “most”, “much”, “named”, “on”, “one”, “other”, “of”, “such”, “that”, “the”, “their”, “these”, “those”, “two”, “while”, “:”. We will call this list the List of Obligatory Left Contexts. The list was built by hand in an iterative fashion, starting by obvious elements and increasing it with other words that we found in the context of valid terms. At this point we only considered single words as they allow very fast context checking procedures.

We repeated the previous experiments now including this additional restriction. Starting by n-grams of size 2, the extraction process generated 18741 candidates (about 35% of the number of

candidates found in the previous experiment). We performed manual checking over the top 30, top 50 and top 100 most frequent n-grams, considering the same three categories as in the previous experiment. The results are shown in table 3:

	Class A	Class B	Class C	P (A only)	P (A + B)
Top 30	28	1	1	90.0%	96.7 %
Top 50	44	3	3	88.0 %	94.0 %
Top 100	87	4	9	87.0 %	91.0 %

Table 3 - The precision of the extraction method using additional contextual filtering.

Using this method, all precision values have increased significantly. Notably, the most relevant improvement has occurred by eliminating most of the incomplete candidates consisting of the last words of a valid TU (those that were related with problem 2). The incorrect candidates still found were:

1. Class B: “antecedent candidates”, “tycho brahe”, “definite description”, “possible antecedents”.

2. Class C: “0cm 0cm”, “ele próprio”, “present paper”, “text temporal”, “system will”, “future work”, “total number”, “uninterrupted linear”, ”fourth binding”.

As we can see, there are still problems such as very frequent genre related constructions (e.g.: “in the present paper” or “the total number of”) that could be tracked down and eliminated, but we found these results good enough for semi-automatic acquisition. It also makes it possible to obtain reasonable candidate lists for n-grams of size 1, because all these restrictions eliminate a great deal of noise. However, for this particular case, precision is still average, because the candidates presented, although relevant words for the domain, are, most of the times, incomplete multi-word TUs instead of complete single-word TUs. Nevertheless, it is sufficient to help the users rapidly collect valid TUs. Again, similar results were obtained for Portuguese, Spanish, French and Italian.

This extraction procedure has been installed and has been in use in *Corpógrafo* for about a year and a half. To avoid affecting the recall of the overall terminological acquisition process, we allow users to choose between three possible filtering options: i) no filtering at all, ii) n-gram filtering using restrictions over the internal form of the candidates, i.e. using just three exclusion lists, and (iii) n-gram filtering on both the internal form and the context, as we have just presented. *Corpógrafo* users have considered these options appropriate for their terminological work.

Future Work

At this point we believe we have found a method that can be improved for a fully automated environment. But for this to be possible, we need to work on some problematic situations. The major problem with the previous approaches is that, because they work on the basis of fixed size n-grams, they are very prone to obtaining incomplete candidates. We are currently re-implementing the algorithm to deal with this problem. We have also already identified problems arising from certain very frequent expressions such as “for example” or “as well as”. We are now in the process of compiling a small lexicon containing this kind of expressions, so that it will be possible in the near future to include filtering capabilities that will reduce the interference of these expressions in the candidate list.

Conclusion

We have presented a simple and robust method to extract terminology from free text. We have explained how this method was developed from its very beginning, describing the problems found along its multiple versions and the solutions found to overcome those problems. The current version of our method uses a set of lexical filters that remove invalid candidates based on certain

context conditions and on the occurrence of certain words inside the candidate n-grams found. Currently, we are able to process technical domain text in Portuguese, English, Spanish, Italian and French with a precision around 90%. The method may be ported to other languages and is now already in use in *Corpógrafo*, a web tool for computational linguistics. The method is also extremely efficient from a computational point of view, allowing performances of two million words per minute on a regular PC. We have also explained how the method may be improved by considering variable length-grams and we described the current work being done in compiling a small lexicon of very frequent fixed expressions in scientific domains, which will enable us to further improve the precision level we now obtain.

Acknowledgements

This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

REFERENCES

BARONI, M. and BERNARDINI, S. (2004): "BootCaT: Bootstrapping corpora and terms from the web" in LINO, M. T., XAVIER, M. F., FERREIRA, F., COSTA, R. and SILVA, R. (eds.): *Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation (Lisboa, Portugal, 25 May 2004)*, p. 1313-1316.
MAIA, B. (2005): "Terminology and Translation – bringing research and professional training together through technology", forthcoming.
MERKEL, M. and ANDERSSON, M. (2000): "Knowledge-lite extraction of multiword units with language filters and entropy thresholds", in *Proceedings of RIAO'2000, Collége de France, Paris, France, April 12-14, 2000*, Vol.1, p. 737-746.
OLIVEIRA, D., SARMENTO, L., MAIA, B. and SANTOS, D. (2005): "Corpus Analysis for Indexing: when corpus-based terminology makes a difference", in the *Proceedings of Corpus Linguistics 2005, Birmingham*.
PASCA, M. (2004): "Acquisition of Categorized named Entities for Web Search", in GROSSMAN, D., GRAVANO, L., ZHAI, C., HERZOG, O. and EVANS, D. A. (eds.): *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management,* p. 137-145.
SARMENTO, L., MAIA, B. and SANTOS, D. (2004): "The Corpógrafo - a Web-based environment for corpora research", in XAVIER, M. F., FERREIRA, F., COSTA, R. and SILVA, R. (eds.): *Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation (Lisboa, Portugal, 25 May 2004)*, p. 449-452.
SEKINE, S. (2004): "Named Entity: History and Future", in <http://cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf> (July 2005).