Term extraction using non-technical corpora as a point of leverage[1]

PATRICK DROUIN

This paper describes a new hybrid term extraction technique for technical corpora. Our main goal is to reduce the amount of noise in the list of candidate terms by restricting the lexical items that can appear inside candidate terms. In order to do so, we base our term extraction process on lexical items selected by a statistical test that targets items that are highly specific to the technical corpus being analyzed.

Keywords: terminology extraction, corpus linguistics, quantitative linguistics, knowledge acquisition.

1. Introduction

Term extraction has been and still is a challenge for anyone interested in domain-specific information retrieval. In the last 10 years, many papers on the subject have been published. Good reviews of the state of the art were recently published in Jacquemin (2001) and Bourigault *et al.* (2001). We can loosely categorize the different techniques as linguistic (David and Plante 1990; Bourigault 1992; Voutilainen 1993; Jacquemin 1996), statistical (Enguehard *et al.* 1992) or hybrid (Daille 1993; Justeson and Katz 1993).

This paper describes a new hybrid term extraction technique for technical corpora that leverages information from non-technical corpora. Our main goal is to reduce the amount of noise in the list of candidate terms (CTs) by restricting the lexical items that can appear inside candidate terms. In doing so, we favor precision over recall[2]. Our research has led to the development of TermoStat, a software tool dedicated to term extraction. The system identifies not only complex terms, but also simple terms, which often tend to be ignored by automated systems.

We first present a technique that aims to extract corpus specific lexical items. These items are identified using a statistical technique that compares frequencies in a technical and a non-technical corpus. The second part of this paper describes how these entries are used as the centerpiece of the term-extraction process. Since our method is corpus-based, we provide detailed information on the nature of the documents included in the corpora. The results and the validation used are then described. Lastly, results obtained with the proposed technique are discussed in terms of precision and recall.

2. Corpus-specific vocabulary

Our work on corpus-specific vocabulary is based on the assumption that a technical corpus contains a set of lexical items that are closely related to its subject and subject area (Kittredge 1982; Picht 1987). We believe that this set can help us gain access to terminological units contained in the technical corpus. The members of this set are those lexical items whose frequency differs significantly from what is considered to be "normal" based on a comparison with previous observations.

Previous work on lexicon specificity has been done by Muller (1979), Lafon (1980), Lebart and Salem (1994) and Camlong (1996) among others. Traditionally, the specific lexicon was identified based on a comparison of the frequency of units in a subcorpus compared to their frequency in the whole corpus. The outcome of the statistical test can lead to three scenarios: the observed frequency can be consistent with the theoretical frequency (SP0); it can significantly higher (SP+) or lower (SP-). We strongly believe that focusing on the context surrounding the lexical items that adopt a highly specific behavior (SP+, SP-) can help us identify terms.

As we mentioned, previous work on the corpus-specific lexicon involved comparison of frequencies observed in a subcorpus to the ones found in the whole corpus. We proposed to tackle the problem differently by using a virtual corpus, called the *global corpus* (*GC*), built at run time from a *reference corpus* (RC) and an *analysis corpus* (AC). The reference corpus is a non-technical corpus while the analysis corpus is a domain-specific, technical corpus. Using two corpora that are quite different in terms of content, we can compare the behavior of lexical units in different types of corpora inside the dynamically built global corpus, and identify the lexical items that are specific to the AC.

Our RC is made up of 13 746 articles taken from *The Gazette*, a Montreal-based newspaper. These articles total up to 7 400 000 words which correspond to approximately 82 700 word forms. The size of the RC, although still modest, can guarantee that the articles cover a wide range of subjects and that their content is heterogeneous. In contrast, as previously mentioned, the AC is domain-specific and topic-oriented. So as to be able to test our algorithms on multiple corpora, we use three different ACs, referred to as $AC_{1...3}$. Table 1 gives detailed information on the size of each corpus.

Table 1. *Description of the corpora*

Corpus	Number of words	Number of word forms
AC_1	11 947	1 207
AC_2	28 583	2 066
AC_3	8 676	1 053

The documents that make up the ACs have been made available to us by the technical writing group of *Nortel Networks*. Assigning a specific domain to a corpus, although often simplistic, is also often quite useful (Kageura 1999). It the case of the ACs, we can say that the topic common to all of the documents is telecommunications, although some concepts are borrowed from the fields of physics, optics and computer science.

It is our intention to use the dichotomy observed between the RC and the AC at the topic level to acquire domain-specific lexical units. In doing so, we are leveraging information about the behavior of words in a non-technical corpus to gain access to terms in a technical corpus. Our work is based on the methods described in Lebart and Salem (1994). This technique, which relies on the standard normal distribution, gives us access to two criteria to quantify the specificity of the items in the set: (1) the *test-value*, which is a standardized view of the frequency of the lexical units, and (2) the *probability* of observing an item with a frequency equal to or higher than the one observed in the AC. Because the probability values decline rapidly, we decided to use the test-value since it permits much more granularity in the results.

In this research, we are interested only in items that appear in the analysis corpus more often than can be predicted from the global corpus (SP+). To isolate them amongst the set of specific items, we use a test-value threshold of +3.09, which means that probability of finding the observed frequency is less than 1/1 000. In other words, the SP- items have a test-value lower than or equal to -3.09, while the SP+ items have a test-value equal or higher to +3.09 and the SP0 items obtain a test-value between these two thresholds. We decided to select this threshold as it gives us a good idea of the importance of a lexical item in the corpus being analyzed. Since our final goal is to extract terms, we will also impose a further constraint on the part of speech and only allow our algorithms to retrieve only nouns and adjectives. We call the elements of this subset of the corpus-specific vocabulary *specialized lexical pivots* (*SLPs*).

The list of SLPs retrieved by TermoStat has been manually reviewed by three professional terminologists working in the field of telecommunications. Table 2 presents the level of precision obtained by TermoStat for SLPs in the three analysis corpora. The terminologists were told to consider an SLP as valid if the item was representative of the domain or the main topic of the corpus. Thus, this is not an evaluation of the precision from a terminological point of view but rather of the relevance of the SLPs.

Table 2. *Evaluation of SLPs*

	AC_1	AC_2	AC_3
Good SLPs	444	810	273
Bad SLPs	84	131	101
Precision	84,1%	86,1%	73,0%

The lexical items that did not classify as SLPs because they did not reach the test-value threshold were not evaluated by the terminologists. Therefore, it is possible that some valid items were not identified during the process. A quick look at the lists of items excluded from the results allows us to verify that this is indeed the case for words such as *time, rate* and *process* in AC$_1$, *house, loss* and *exchange* in AC$_2$ and *point, building, state* and *manager* in AC$_3$.

The specificity of these words cannot be observed strictly from a lexical point of view; one must also look at semantic aspects. Multiple phenomena come into play in this case, including homonymy, polysemy and de-terminologization (Galisson 1978; Meyer and Mackintosh 2000). Without an additional level of tagging that could take meaning into account, these items cannot be accurately retrieved by the specificity test.

One might also wonder what the influence of the RC is on the results. In order to evaluate its impact on the list of SLPs, we divided the RC into four different sections of equal size (called RC$_{1...4}$) and created three reference corpora named RC$_2$, RC$_4$ and RC$_{1+3}$. The first two correspond to one section of the original corpus while the last corresponds to the concatenation of two sections. Using these corpora, we generated a new list of SPLs from AC$_1$ and compared the results obtained. This comparison allows us to test the stability of the list of SLPs produced in relation to the RC used.

Our first step was to compare results obtained with corpora of equal size (RC$_2$ and RC$_4$). The second step was to evaluate the results obtained using corpora of different sizes (RC versus RC$_{1+3}$ and RC versus RC$_4$). The comparison of RC to RC$_4$ maximizes the size difference between the reference corpora being used to retrieve SLPs.

We manually identified the number of differences in the lists of SLPs generated by TermoStat. Figure 1 shows the number of differences noted between the lists for the various frequency groups. For example, for SPLs that have a frequency of 1, we identified 27 differences between the lists generated using RC and RC$_{1+3}$.

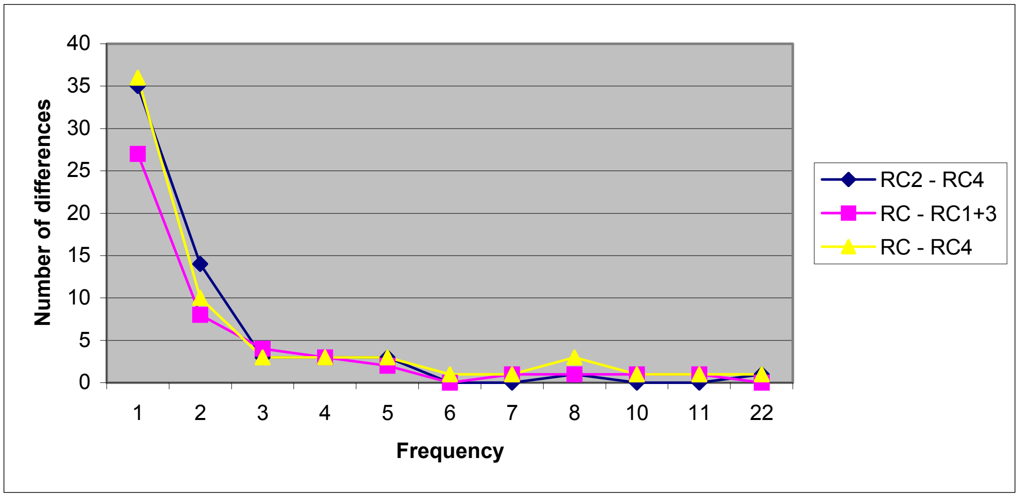

Figure 1. *Results of the stability test for SLPs*

As we can see in Figure 1, although the observed behavior is generally similar for all the tests, we observed a variation for the lower frequency groups. The total discrepancy between the various lists is not very significant and is in the order of 20% for RC and RC_4, 11.5% for RC_2 and RC_4 and 9.2 % for RC and RC_{1+3}.

Labbé and Labbé (2001) came to a similar conclusion and demonstrated that the test used to measure the specificity of a lexical item should not be systematically trusted for frequencies lower than 2. From a statistical point of view, this makes seems logical. From a terminological point of view, we believe that although the results are not statistically significant, they can be very useful. Although we note that the RC used to extract the SLPs has an influence on the results of the test, the terminologists who reviewed our data confirmed their relevance.

3. Term extraction

This section of the paper describes the term extraction technique that has been implemented in the prototype called TermoStat. In order to illustrate the algorithms, the process we use and the effects of constraints we impose during the extraction process, we will be using the following example taken from one of the ACs.

> For/IN Dual/JJ MSA/NNP sites/NNS (/(line/NN sites/NNS with/IN high/JJ OADM/NNP counts/NNS)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN flow/NN is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ MSA/NNP (/(DSCM/NNP or/CC OADM/NNP filter/NN)/SYM is/VBZ placed/VBN between/IN the/DT Booster18/NNP and/CC Booster21/NNP circuit/NN packs/NNS ./.

TermoStat uses corpora previously tagged with Eric Brill's tagger (Brill 1994). The first step taken by the software is to locate all headwords. We consider that any noun that belongs to the list of SLPs may be considered as a headword. These headwords are used as the starting point of the term extraction process that analyzes the corpus from right to left. If we take the previously used example and identify headwords, we get the following output:

> For/IN Dual/JJ MSA/NNP sites/NNS (/(line/NN sites/NNS with/IN high/JJ OADM/NNP counts/NNS)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN flow/NN is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ MSA/NNP (/(DSCM/NNP or/CC OADM/NNP filter/NN)/SYM is/VBZ placed/VBN between/IN the/DT Booster18/NNP and/CC Booster21/NNP circuit/NN packs/NNS ./.

Our algorithm revolves around the concept of boundaries introduced by Bourigault (1992). We consider a CT to be delimited by these boundaries[3]. Our view of the concept is a little different that the one originally defined by Bourigault (1992), as we take into account not only the part of speech of the lexical units being considered, but also the results of our statistical process and some parts of the formatting of the corpus. In order not to qualify as boundaries, the words must appear in the list of SLPs.

As we defined them, the boundaries are thus context-independent and do not rely on local syntactic analysis. The underlined units in the following example are considered to be boundaries and will delimit the CTs.

> For/IN Dual/JJ MSA/NNP sites/NNS (/(line/NN sites/NNS with/IN high/JJ OADM/NNP counts/NNS)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN flow/NN is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ MSA/NNP (/(DSCM/NNP or/CC OADM/NNP filter/NN)/SYM is/VBZ placed/VBN between/IN the/DT Booster18/NNP and/CC Booster21/NNP circuit/NN packs/NNS ./.

Previous term extraction studies suggested that the length of candidates retrieved by a system should be limited. Table 3 presents the breakdown of the observations done by Justeson and Katz (1993) and Nkwenti-Azeh (1994).

Table 3. *Previous studies on the length of terms and CTs*

	1 word	2 words	3 words	4 words	5+ words
Justeson and Katz	29.5%	54.5%	12.4%	3.9%	0%
Nkwenti-Azeh (1)	9.15%	71.86%	16.93%	2.06%	0%
Nkwenti-Azeh (2)	7.30%	49.02%	32.83%	8.88%	1.97%
Nkwenti-Azeh (3)	30.73%	49.84%	15.13%	3.5%	0.8%

Although the data used by Justeson and Katz (1993) were taken from dictionaries, tests that they performed on corpora led to similar results. In contrast, Nkwenti-Azeh (1994) used data taken from various sources of information including a corpus (1), a terminology database (2) and a technical dictionary (3).

Justeson and Katz (1993) indicate that, overall, as the length of the candidate increases, the likelihood of it being a valid terminological unit decreases. However, they did note that terms with a length of 2 words are generally more frequent than single-word terms. Moreover, they also indicated that corpora related to the medical domain usually constitute an exception, and the reverse distribution is observed. That leads us to believe that a term extraction tool cannot systematically exclude single-word terms because they could represent key concepts for certain domains.

Table 3 suggests that using a window of a maximum of 6 words is sufficient to cover most cases of complex terminological units. Although this can lead to the exclusion of some terms from the final list of CTs, as we stated before, we prefer to limit the recall of the system in order to increase its precision. The potential structure of CTs that will be retrieved by TermoStat can be described using the following regular expression:

> (A|N)? (A|N)? (A|N)? (A|N)? (A|N)? N
>
> where:

- A is an adjective,
- N is a noun,
- (A|N) is a noun or an adjective,
- ? represents zero or one occurrence of the element,
- ___ is an element that belongs to the SLP set.

This grammar is similar to the one used by Frantzi and Ananiadou (1997) and Justeson and Katz (1993). Our grammar differs from the ones used by these authors in that it does not allow for the inclusion of elements other than SLPs inside the CTs and that it imposes a length constraint. We can now look at the example we used before and illustrate the impact of choosing a window of six words on the term extraction process.

> For/IN Dual/JJ [[[MSA/NNP] [sites/NNS]] (/([[line/NN] [sites/NNS]] with/IN high/JJ [[OADM/NNP] [counts/NNS]])/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN [flow/NN] is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ [MSA/NNP] (/([DSCM/NNP] or/CC [[OADM/NNP] [filter/NN]])/SYM is/VBZ placed/VBN between/IN the/DT [Booster18/NNP] and/CC [[Booster21/NNP] [[circuit/NN] [packs/NNS]]] ./.

The units between brackets make up the first list of CTs. This list will be filtered once by TermoStat in an attempt to verify the autonomy of each of them in the AC so as to eliminate term fragments. We consider any candidate (CT_1) included in a longer candidate (CT_2) and never appearing outside of CT_2 to be a term fragment. Term fragments are excluded from our final results list. For example, in the previous example, *booster21*, *booster21 circuit*, *circuit* and *circuit packs* would be eliminated since they do not occur outside of *booster21 circuit packs*.

4. Defining the gold standard

In order to evaluate the quality of the output of TermoStat accurately, the results were submitted to a two-stage validation process. The first step is an automated validation, while the second one relies on human interaction. This whole process is not part of the actual TermoStat software but was built in solely for the evaluation of its performance.

The automated validation consists of a comparison of the CTs with a list of terms found in a terminology database. As is the case with the AC, the database was provided to us by *Nortel Networks*. The automated process relies on a comparison of strings and does not include any sophisticated analysis to determine if two matching strings have the same meaning. This approach is similar to the one described in Daille (1994). We believe that the close relationship between the terminology stored in the database and the ACs is enough to avoid hiccups during the validation process.

As is to be expected, the comparison of the CT list and the content of the terminology database did not lead to the validation of all the CTs. The remainder of the list was submitted to a team of three terminologists. These terminologists all have a very good knowledge of the telecommunications industry. Each member of the review team was given a CT list linked to one of our three ACs. After the initial review, the most senior terminologist reviewed the results one more time for the three corpora.

As with any validation process that relies on humans, we cannot assert that the results obtained are free of subjectivity. We strongly believe that different terminologists will identify different terms in the same document and that the same phenomenon could be observed with one terminologist looking at the same corpus over a period of time. It would be very interesting to be able to take into account the human influence over the validation process but doing so is beyond the scope of this paper.

The process we described in the previous paragraphs will be useful when trying to evaluate the precision of the results, since every CT retrieved by TermoStat was validated. As far as recall is concerned, the usual procedure is to have someone go through a corpus and identify every term it contains. Unfortunately, we could not undertake such a task so we had to look for a different solution. Since our goal is to determine the impact of the use of SLPs on the term extraction, we decided to modify TermoStat so as not to use the SLP constraint when proceeding to term extraction and to use the results obtained as our reference list for the evaluation of recall. This modification of the software was for the sole purpose of evaluating the performance of TermoStat and is not part of the system in itself.

5. Results

In this section, results obtained using TermoStat are described using the standard measures of precision and recall. The frequency threshold used in the following figures corresponds to the minimal frequency a CT must reach to be retained. In other words, the CTs that qualify all have a frequency equal to or higher than the threshold specified.

5.1 Recall

Any term extraction process that applies restrictions on lexical items that can appear inside CTs leads to some loss of information. In this section, we try to quantify the impact of our technique on the level of recall obtained by TermoStat. As one can observe in Figure 2, the performance of the software increases rapidly with the frequency threshold.

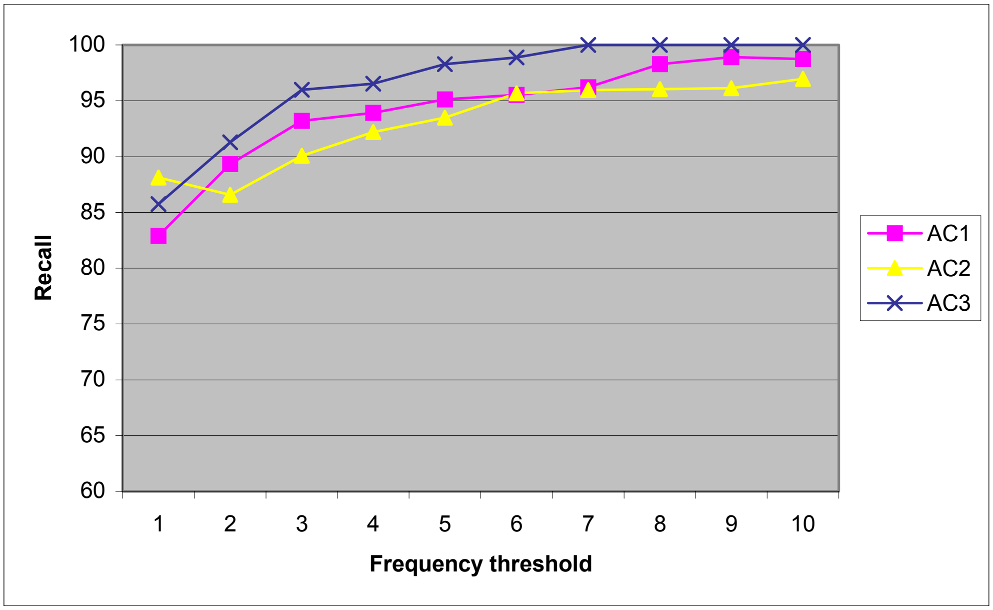

Figure 2. *Overall recall*

Although this could be considered to be contradictory, it can be explained by the method we use to evaluate recall, as described in the previous section. We do not compute recall on the total number of terms found in the corpus but rather on the number of CTs retrieved by TermoStat when used without the SLP constraint. The frequency threshold indicates the minimum frequency that a CT must meet in order to be retained by the system. That is to say, for example, that when looking at CTs that have a frequency higher than or equal to 2, use of the SLP constraint resulted in the extraction of 86.6% of the CTs identified without the constraint. As we can see from this figure, the SLP constraint does not have much of an impact for frequencies higher than 3.

This filtering of data and the emphasis on precision over recall is in line with our original goal, to reduce the amount of noise in the CT list regardless of any negative effect this might have on the coverage of our algorithm. As we can see from Figure 2, this impact is not drastic.

5.2 Precision

The overall precision reached by the two prototypes is very similar and, at first glance, is not very impressive. Figure 3 displays the results obtained by TermoStat for the three corpora with (AC_1+, AC_2+, AC_3+) and without (AC_1, AC_2, AC_3) the SLP constraint. We notice a slight difference in the lower frequency range but the trend tends to disappear as the frequency threshold increases. It is interesting to note that, overall, the constraint does not lead to a loss of precision but rather to a gain of 2.44 %.

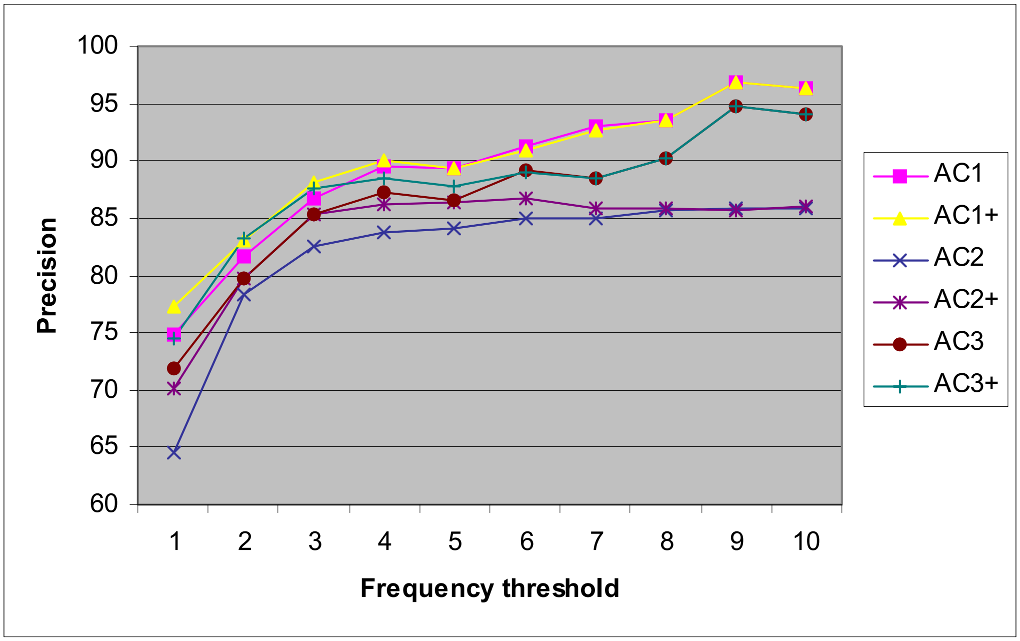

Figure 3. *Overall precision*

As expected, the quality of the results, in terms of precision, increases as the frequency threshold also increases. This confirms earlier claims that the frequency is a very good indicator of the quality of a CT (Daille 1994).

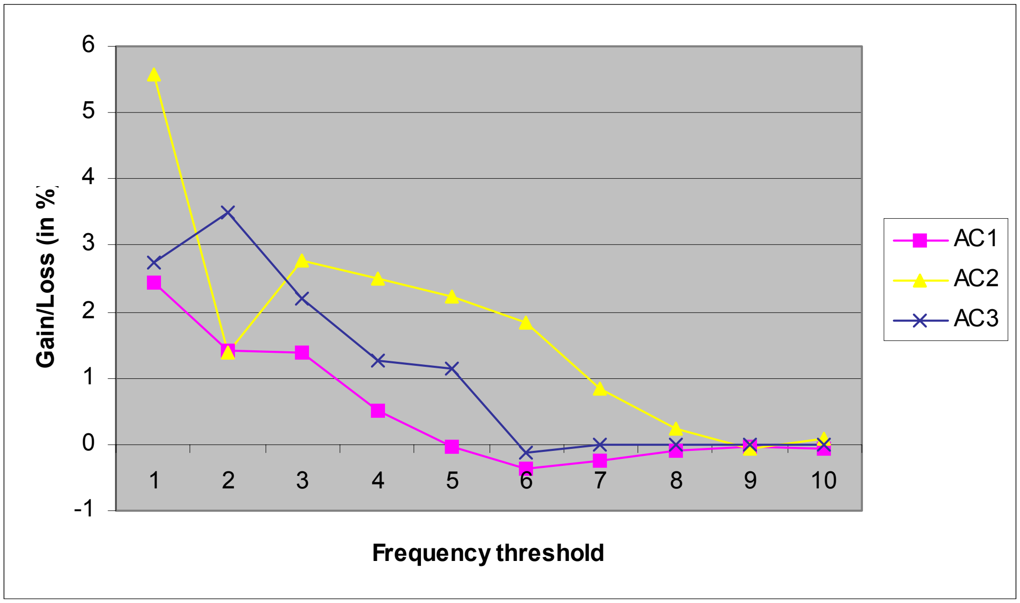

Figure 4. *Overall precision variation*

Figure 4 gives a better idea of the variation of precision observed when enforcing the SLP constraint during the term extraction process. For frequencies lower than 3, we notice that it leads to an increase in precision. One can also observe that, for frequencies higher than 3, the proposed technique leads to a decrease in precision. It is also quite obvious that the benefits of the SLP algorithm disappear as the frequency threshold increases.

We can get a better understanding of the performance of the algorithms when we divide the results into simple candidate terms and complex candidate terms. As Figures 5 and 6 show, depending on which group we look at, the variation of precision is drastically different. For complex candidate terms, the pattern is similar to the one observed for the overall results and the precision gain is less significant below a frequency threshold of 3. In the case of AC_1 and AC_3, we can even see that the SLP constraint does not lead to any variation in precision for higher frequencies.

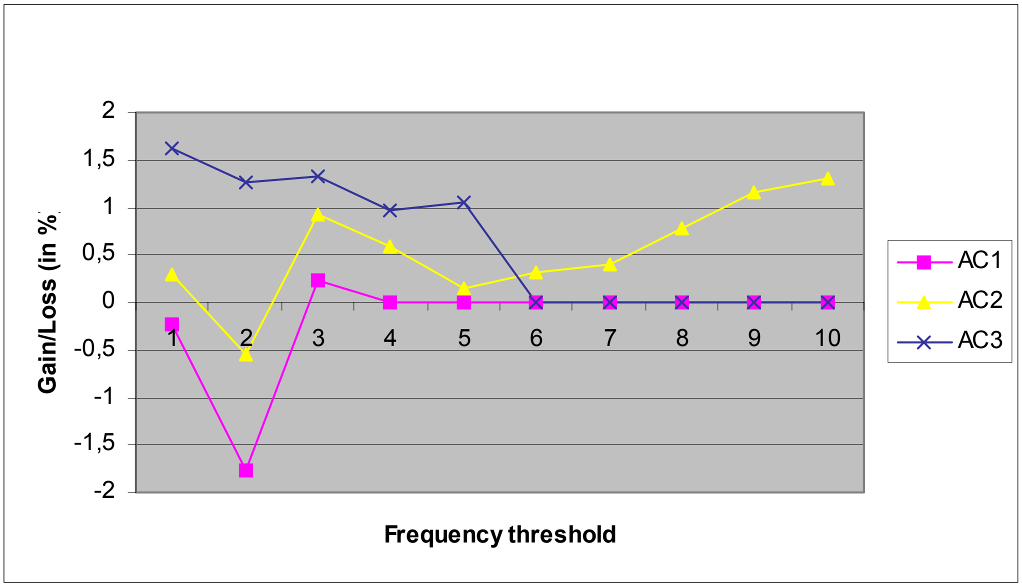

Figure 5. *Precision variation - Complex CTs*

Even though we see a slight gain in the case of AC_2 and no significant loss for the other two corpora, Figure 5 leads us to conclude that a term extraction process that relies on SLPs does not lead to added value when looking at complex terms. This conclusion should come as no surprise since the original work done to pinpoint the corpus-specific lexicon was not aimed at compound items but at words.

Although we believed that SLPs would allow us to gain access to longer units, it seems that a technique that relies on boundaries might not capture the complexity of the phenomena that come into play in compound terms. A different technique that would rely on a second level of statistical analysis such as mutual information (Church and Hanks 1989) could be beneficial. Until we test this hypothesis, the SLP approach can still be used since it leads to an increase of precision overall and even in cases when it does not lead to some gain in precision, the loss is very small.

When we take a closer look at the precision obtained for single word CTs (Figure 6), we can see that the performance of TermoStat is very interesting. As was the case for the overall results and the complex CTs, the benefits of the SLPs decrease as the frequency gets higher. The boost in performance seen for lower frequencies is quite surprising. Simple CTs are often ignored by term extraction software since their recognition is problematic. The same thing can be said for the low-frequency CTs, even for complex CTs. When we consider these facts, the benefits of the SLP technique, which makes it possible to pinpoint low-frequency single CTs, become obvious. Being able to retrieve these candidates is also interesting for human terminologists who struggle, even with concordance software or other tools, to find them in large bodies of running text.

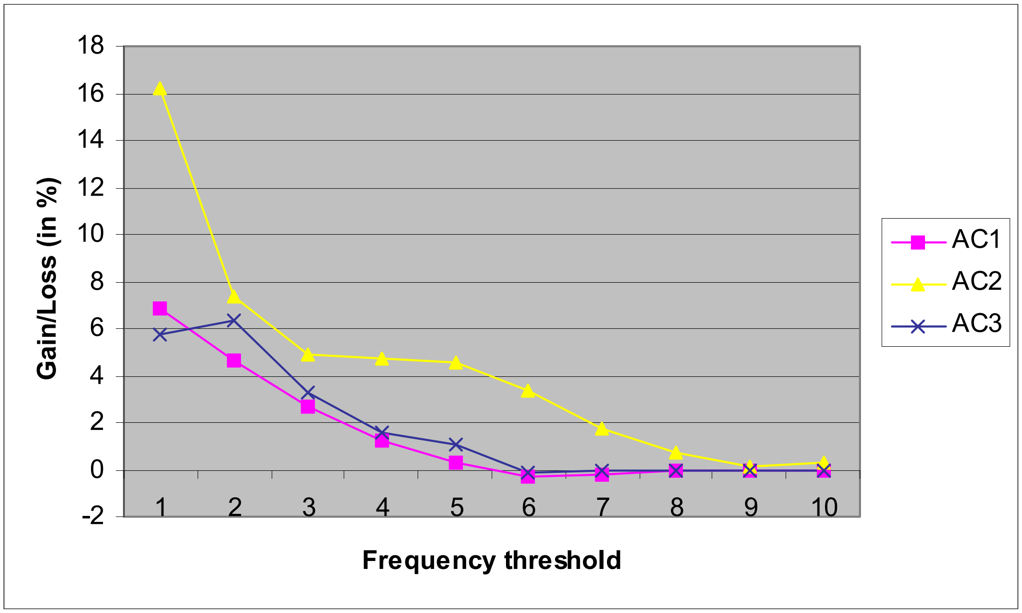

Figure 6. *Precision variation - Simple CTs*

As stated before, the frequency of a CT is useful in determining whether we are looking at a term or not, but our results seem to indicate that it is even more interesting to verify if the frequency observed is the one that we could expect to see based on a larger non-technical corpus. We can then envision a term extraction process that would take advantage of the strengths of both clues and use the SLP technique for single word items and a raw frequency threshold for complex terms.

One more conclusion can be drawn from our work: although Labbé and Labbé (2001) proved that the specificity test is unstable when looking at lower frequency items, we can still use its results for term extraction and obtain good quality CTs. The test might lead to results that are not interesting from a statistical point of view, but are highly useful from a terminological standpoint.

6. Concluding remarks

The first step of the term extraction approach proposed in this paper relies on the recognition of a corpus-specific lexicon as originally proposed by Muller (1979). The technique has been modified so as to leverage information from a non-technical corpus in order to evaluate the specificity of lexical items in a technical corpus. We proposed to restrict our use of the specific items and apply a few constraints on the nature of the items retrieved: they must be extracted from a technical corpus, they must have a frequency in this last corpus which is significantly higher than expected (less than 1/1 000 as compared to a non-technical corpus), and they have to be either nouns or adjectives. We call this last set of items *specialized lexical pivots* or *SLP*s. The overall precision obtained in the set of SLPs retrieved by the TermoStat was 81%.

The corpus-specific vocabulary previously identified is then used as the starting point of our term extraction process. The technique we describe leads to very good results when applied to the extraction of simple terms, reaching an overall precision of about 86% for this subset (74% overall). SLPs are particularly interesting when trying to identify low-frequency terms. The only minor yet notable drawback of the SLP technique for simple-term extraction is the loss of some items that seem to have a semantic specificity rather than a purely lexical specificity. We believe that further work and corpus-based investigations on de-terminologization (Galisson 1978; Meyer and Mackintosh 2000) would help tackle this problem.

Even though TermoStat showed good performance when extracting complex terms (65%), further work remains to be done, since the performance level reached by TermoStat with these terms is not as good as that observed for simple terms. One potential solution would be to use a second level of statistical information such as mutual information (Daille 1994) or some termhood-weighting factor (Frantzi and Ananiadou 1997; Nakagawa and Mori 2002) applied to the extraction process itself instead of a linguistics-based extraction.

As with any corpus-based research, we need to extend the scope of our work so as to validate it against other corpora. In doing so, we would like to broaden our research to corpora from various domains and maximize the size of the reference corpus to ensure the stability of the SLPs. The results obtained so far are encouraging, but further corpus-based investigations are mandatory.

Notes

1. The author would like to thank Andy Lauriston, Lynne Leslie and Tricia Morgan for the time they spent reviewing the results and approving the CTs. I would also like to thank Elizabeth Marshman who kindly accepted to review this paper.

2. Precision is the proportion of retrieved items that are relevant; recall is the proportion of all the relevant items that is retrieved.
3. The term extraction process of TermoStat has since been rewritten using a finite state automaton.

References

Bourigault, D. 1992. "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases". *Proceedings of the Fourteenth International Conference on Computational Linguistics-COLING 92.* Nantes, 977-981.

Bourigault D., C. Jacquemin and M.-C. L'Homme (eds). 2001. *Recent Advances in Computational Terminology.* Amsterdam/Philadelphia: John Benjamins.

Brill, E. 1994. "Some advances in transformation-based part-of-speech tagging". *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*. Seattle, 722-727.

Camlong, A. 1996. *Méthode d'analyse lexicale textuelle et discursive.* Paris: Orphrys.

Church, K. W. and P. Hanks (1989). "Word association norms, mutual information and lexicography". *Computational Linguistics* 16(1), 22-29.

Daille, B. 1993. "Extraction automatique de terminologie monolingue". *Actes du colloque Informatique et langue naturelle*. Nantes.

Daille, B. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Paris: Université de Paris 7.

David, S. and P. Plante. 1990. "De la nécessité d'une approche morpho-syntaxique en analyse de textes". *Intelligence artificielle et sciences cognitives au Québec* 2(3), 140-155.

Enguehard, C., P. Malvache and P. Trigano. 1992. "Indexation de textes : l'apprentissage automatique de concepts". *Actes du XVème colloque international en linguistique informatique*. Nantes, 1197-1202.

Frantzi, K. T. and S. Ananiadou 1997. "Automatic Term Recognition Using Contextual Cues". *Proceedings of the 3rd DELOS Workshop*. Zurich, offprint.

Galisson, R. 1978. *Recherches de lexicologie descriptive : la banalisation lexicale*. Paris: Nathan.

Jacquemin, C. 1996. "What is the tree that we see through the window: A linguistic approach to windowing and term variation". *Information Processing & Management* 32(4), 445-458.

Jacquemin, C. 2001. *Spotting and discovering terms through natural language processing*. Cambridge: MIT Press.

Justeson, J. and S. Katz 1993. *Technical terminology: some linguistic properties and an algorithm for identification in text. Technical Report RC 18906*. IBM Research Division, offprint.

Kageura, K. 1999. "Theories 'of' terminology: a quest for a framework for a study of term formation". *Terminology* 5(1), 21-40.

Kittredge, R. 1982. "Variation and Homogeneity of Sublanguages". In R. Kittredge and J. Lehrberger (eds). *Sublanguage : Studies of Language in Restricted Domains*. Berlin/New York: de Gruyter, 107-137.

Labbé, C. and D. Labbé 2001. "Que mesure la spécificité du vocabulaire ?". *Lexicometria* 3.

Lafon, P. 1980. "Sur la variabilité de la fréquence des formes dans un corpus". *MOTS* 1, 128-165.

Lebart, L. and A. Salem 1994. *Statistique textuelle.* Paris: Dunod.

Meyer, I. and K. Mackintosh. 2000. "When terms move into our everyday lives: An overview of de-terminologization". *Terminology* 6(1), 111-138.

Muller, C. 1979. *Langue française et linguistique quantitative : recueils d'articles.* Genève: Slatkine.

Nakagawa, H. and T. Mori 2002. "A simple but powerful term extraction method". *Proceedings of the 2nd International Workshop on Computational Terminology (COMPUTERM 2002).* Taipei, 29-35.

Nkwenti-Azeh, B. 1994. "Positional and combinational chacteristics of terms: consequences for corpus-based terminography". *Terminology*. 1(1), 61-95.

Picht, H. 1987. "Terms and their LSP Environment — LSP Phraseology". *Meta* 32(2), 149-155.

Voutilainen, A. 1993. "Nptool, a detector of English noun phrases". *Proceedings of the Workshop on Very Large Corpora.* Columbus, 48-57.

Term extraction using non-technical corpora as a point of leverage

Abstract

This paper describes a new hybrid term extraction technique for technical corpora. Our main goal is to reduce the amount of noise in the list of candidate terms by restricting the lexical items that can appear inside candidate terms. In order to do so, we base our term extraction process on lexical items selected by a statistical test that targets items that are highly specific to the technical corpus being analyzed.

Keywords: terminology extraction, corpus linguistics, quantitative linguistics, knowledge acquisition.

Biographical note

Patrick Drouin has recently been appointed professor in the translation and linguistics department of the Université de Montréal, where he teaches localization. He obtained his Ph.D. in linguistics from the Université de Montréal in 2002, working on term extraction using hybrid techniques. He also worked in the private sector as a translation and terminology technology specialist from 1996 to 2002.

Patrick Drouin
Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), H3C 3J7
patrick.drouin@umontreal.ca