Towards Automatic Extraction of Monolingual and Bilingual Terminology

Béatrice DAILLE* - Éric GAUSSIER - Jean-Marc LANGÉ
(*)TALANA Univ. Paris 7 & IBM-France, Sce 3099, 75592 Paris Cedex 12
Email: jml@vnet.ibm.com

Abstract

In this paper, we make use of linguistic knowledge to identify certain noun phrases, both in English and French, which are likely to be terms. We then test and compare different statistical scores to select the "good" ones among the candidate terms, and finally propose a statistical method to build correspondences of multi-words units across languages.

Acknowledgement

Most of this work was carried out under project EUROTRA ET-10/63, co-sponsored by the European Economic Community.

1 Introduction

A technical lexicon consists of simple as well as complex lexical units. Complex lexical units are mainly compounds, commonly characterized by a set of morphosyntactic and semantic criteria (see for example [Benveniste, 1966] for French and [Hatcher, 1960] for English). The constitution of a list of technical terms from a specific domain is an important issue in Natural Language Processing, for example in applications such as Machine Translation. Several methods for the extraction of terms have been proposed, relying on linguistic or statistical approaches. To build a monolingual terminology bank, [Bourigault, 1992] has implemented a software package which provides a list of likely terminological units, based on linguistic specifications. Different statistical methods have also been tested to extract collocations from large corpora, as [Church and Hanks, 1990, Smadja and McKeown, 1990]. Our work makes use of both linguistic and statistical knowledge. The outline of this paper is the following: first, given linguistic specifications of English and French terms, we select noun phrases which are likely to be terms, and which we will call *candidate terms*. To build a monolingual terminology bank, we then apply different statistical scores to select the good ones among the candidate terms. Finally, we test statistical methods to obtain a list of bilingual terms, using aligned corpora in French and English. All work and tests are carried out on a parallel bilingual corpus that is tagged by the words' part-of-speech and lemma. This corpus consists of about 200,000 words for each language in the field of telecommunications. Similar work can be found in [Van der Eijk, 1993].

2 Linguistic specifications of French and English terms

2.1 Basic multi-words units

From an investigation of the corpus, and from the terminologists' point of view (for example [Nkwenti-Azeh, 1992]), it appears that most of the terms encountered in technical fields are noun phrases corresponding to a limited number of syntactic patterns. These patterns define the morphosyntactic structures of multi-words units (MWU) which are likely to be terms. We define the length of a MWU as the number of *main items* it contains. A *main item* is either a noun, an adjective, a verb or an adverb; thus, neither determiners nor prepositions are considered *main items*. MWU of length 1 often correspond to words connected with a hyphen or an apostrophe, and are easy to identify. We will not deal with them in this paper. Nor are we concerned with mono-word terms (e.g. *telematics*), the extraction of which is a different -possibly

more complex- problem. We present below the patterns retained for MWU of length 2, which we will refer to as **base MWU**. For each pattern we give an example followed by its translation in parentheses. The interpretation of the abbreviations is straigthforward.

French

- **N Adj**
  orbite géostationnaire
  (geostationary orbit)
- **N1 N2**
  diode tunnel (tunnel diode)
- **N1 de (det) N2**
  bande de fréquence (frequency band)
- **N1 prep (det) N2**
  assignation à la demande
  (demand-assignment)

English

- **Adj N**
  multiple access (accès multiple)
- **N1 N2**
  data transmission
  (transmission de données)

2.2 Operations on base MWU

Although MWU of length greater than 2 do exist, they are most of the time built from base MWU by one of the following operations, encountered both in French and English (the base MWU is bracketed in the examples):

- overcomposition:
  régénération des [lobes latéraux]
  [side lobe] regrowth
- modification:
  [station terrienne] brouilleuse
  interfering [earth(-)station]
- coordination:
  assemblage et désassemblage de paquets
  packet assembly/disassembly

Overcomposition and modification do not always preserve the base-MWU structure: in French, an adjective modifier is often inserted inside a base-MWU of `N1 prep N2` structure; for example, *national (domestic)* can be inserted inside *réseau à satellites (satellite network)* to produce *réseau national à satellites (domestic satellite network)* (notice that in English, the adjective precedes the base MWU most of the time); in English, the `Adj N` structure is altered when two base-MWU, one of `Adj N1` structure and one of `N2 N1` structure, are overcomposed: *geostationnary satellite* and *telecommunication satellite* yield *geostationnary telecommunication satellite*. Coordination, which requires two base-MWU, always breaks the structure of one of them. These operations are not used with the same frequency, composition and modification being more frequent than coordination.

From the preceding considerations, it seems natural to focus on base MWU. But, we take into account the cases where base-MWU structure is broken as well as their variants. Variants of base-MWU are mainly graphical, orthographical and morphosyntactic (for more details see [Daille, 1994]). In English, we have also accepted the transformation of a base-MWU of `N2 N1` structure into a `N1` *of* `N2` structure.

2.3 Extraction of base MWU

To extract and count the occurrences of base MWU (under their various forms) we use their morphosyntactic structures. Happily enough, our corpus is tagged: we can rely on the sequences of part-of-speech to extract the relevant candidates. For this purpose, we have implemented finite automata associated to regular expressions covering most occurrences of morphosyntactic structures, including modifications due to the afore mentionned operations. The output of the program is a list of pairs composed of two lemmas. Under each pair are stored all occurrences of the base MWU extracted from the corpus: for example, for the pair
(interference, level),
we get the following sequences:
interference level, interference levels, level of interference, levels of interference;
for the pair
(circuit, numérique),
we get:
circuit numérique, circuits numériques, circuit entièrement numérique, circuits analogiques et numériques.

Unfortunately, the pairs obtained through this method are not always "good" terms; we therefore have to introduce additional filters: statistical scores which use the number of occurrences of the pairs as input.

3 Statistical scores to select good monolingual candidates

Since the MWU extracted after the linguistic specifications still contain noise, we use statistical scores as an additional filter in order to choose the "good" ones among the candidate base MWU (or candidates for short). These scores apply on *candidate lemma pairs* for a given morphosyntactic pattern. We computed different scores: frequencies, association criteria, Shannon diversity and distance measures. We discuss frequencies and association criteria below. Informations provided by Shannon diversity and distance measures are presented in [Daille, 1994]. Most of the association criteria can be found in the classic literature, such as [Clifford and Stephenson, 1975], and are based on the so-called "contingency tables". In the field of computational linguistics, *mutual information* [Brown *et al.*, 1988], $\phi 2$ [Church and Hanks, 1990], or a likelihood ratio test [Dunning, 1993] are suggested.

Our testing method consists in comparing our result list, sorted according to a specified score, with a reference list containing only valid terms. In order to build the reference list, we augmented an existing terminology database (EURODICAUTOM) with hand work: we selected those candidates for which at least two judges out of three agreed on their goodness, and included them in the reference list. A candidate will be considered "good" when it is found in this reference list.

The program for the extraction of candidates was run on a 240,000 words French corpus, divided into 9,541 sentences. It yielded 2,400 candidates that appear at least three times in the text. After sorting the candidates according to the particular score examined, we build a graphic representation of the proportion of "good" candidates per range of score values, in the form of a histogram. Figure 1 shows the histograms obtained on the French N1 de N2 candidates with two different scores: the likelihood ratio and mutual information. The vertical axis, which represents the proportion of good candidates, is bounded between 0 and 1. But, since we do not pretend to have built an exhaustive reference list, we can't actually expect values of more than 0.8. The horizontal axis grows with the score. The likelihood ratio test (LOG) (the use of a likelihood ratio test leads to a log likelihood statistics, which explains the LOG abbreviation) shows a curve that grows slowly at first, and then decidedly. There is a true opposition in the behaviours of small and big values, the representativity of good candidates approaching 0.8 in that case. On the contrary, mutual information (MI) is quite uniformely distributed, and the representativity never exceeds 0.5. Thus, if one had to choose between these two scores, the first one would undoubtedly be selected.

The outcome of the study of the histograms is that very few scoring methods will meet our objective. The most significant seem to be the likelihood ratio test, and the number of occurrences of the pair. If it were for the sole histograms, one could be tempted to keep only the frequency of the pair, but due to its definition unfrequent terms won't stand out; therefore, we have to keep several methods. Anyhow, the performance of our scoring functions indicate the best method to follow: start any task (human post-editing for example) with the highest frequency values on monolingual candidates previously identified via linguistic patterns, thus maximizing the odds that the candidates will indeed be "good" candidates.

Is the best score a combination of scores? An extra data analysis method

To complete the evaluation and comparison of the different scores used to select actual terms from our candidates list, we have applied to these scores specific data analysis methods. In our case, the techniques of data analysis can provide at least two things: a probabilistic study of the correlations between scores, in order to select the relevant ones, and the possibility of combining scores, in order to improve the results.

We carried out Principal Component Analysis (PCA) with a set of fifteen variables, i.e. the scores, on a set of 1,230 individuals, i.e. the can-

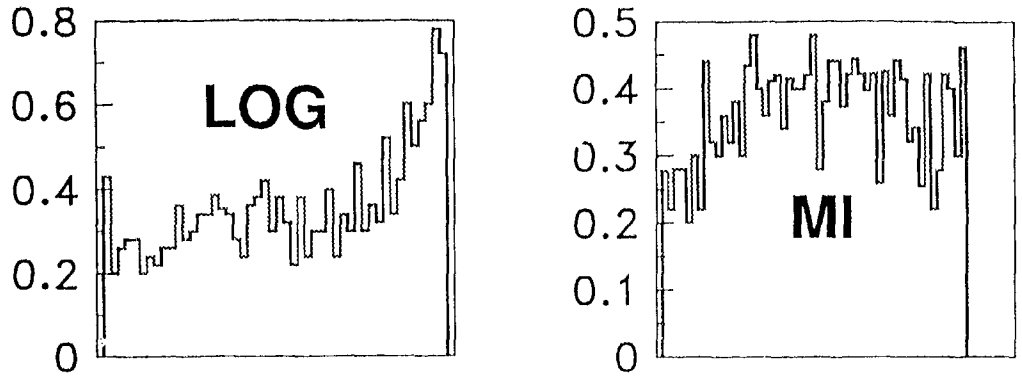

Figure 1: Histograms corresponding to the log likelihood (left), and to mutual information (right)

didates. The first objective of PCA is to reduce the number of variables with as little loss of information as possible. Doing so should enable us to visualize our data more easily. From the study of the correlation matrix of variables and their part in the principal axis, it appeared that variables could be divided into four classes. In each class, we wanted to retain the most significant variable with regard to our task. This was done by selecting in each class the variable with the best score-histogram. All other variables can be confidently connected to one of these.

With the remaining four variables we tried to investigate the set of candidates. The difficulty of this task comes from the high number of candidates, so we decided to work on smaller sets. We divided the whole set of candidates into six subsets depending on the number of occurrences of the candidate (respectively with a number of occurrences in the text equal to 2, 3, 4, 5, between 6 and 10, and over 10). This division was used to counterbalance the fact that most scores tend to privilege low frequencies. We then visualized the candidates graphically (with different tick marks distinguishing terms and non-terms) on several two-dimensional diagrams, the axis of which are linear combinations of the different variables. The results show series of marks overlapping each other, without any clear separation between terms and non-terms.

The PCA allowed us to go further in the study of the statistical scores, and confirmed that only a few are relevant to our task; moreover, we found no satisfactory combination of variables that could account for termhood.

4 Bilingual candidate term alignment

The problem is now one of finding correspondences between candidates across languages (between English and French in our case). As it has been shown in the previous section, it is difficult to restrict the set of candidates to the "good" ones. Furthermore, if a candidate is really a term, it should be easy to find the equivalent in other languages, or more precisely, in our case, in aligned sentences (supposing that terms are always translated into the same lexical unit). For these reasons, we tried to pick out associations between the whole sets of candidates, both French and English. Sticking to our general philosophy of harmonizing linguistic and statistical contributions, we made experiments with two methods.

4.1 A simple count method, and a possible improvement

This method basically counts how often a source candidate and a target candidate occur in aligned sentences; we shall call the resulting score *bilingual count*. For a given source language candidate, the most frequently aligned target candidates are sorted according to the score; one can then apply the criterion defined in [Gaussier *et al.*, 1992], which consists in extracting a bilingual pair only when the target candidate has no stronger association with any other source candidate. Such a naive frequency collection gives satisfactory results, providing a sorted list where the top half candidates are

mostly good. We can however try to improve on that result by taking advantage of available linguistic information, namely the part-of-speech patterns of the candidates. We estimated, by counting pattern alignments on a bilingual corpus, *pattern affinity*, which we define as the probability that the pattern of the source candidate gets translated into the pattern of the target candidate (e.g. terms with pattern NdeN in French get translated to terms with pattern NN in English in 81% of the cases). Now, instead of simple counts of the occurrences of the source and target candidates, we use counts that are weighted with a function of the pattern affinity of candidates, as explained in the following formula:

$$C(s,t) = \Sigma_{sent} \frac{probes(p_t \mid p_s) \times C(t)}{\Sigma_x probes(p_x \mid p_s) \times C(p_x)}$$

where $C()$ means "count of", and p "pattern of"; *probes* stands for the probabilities we estimated, and *sent* represents the set of aligned sentences in which both the source candidate, s, and the target candidate, t, appear. Results are discussed below.

4.2 Another method: using word alignment scores

This method consists essentially in aligning terms using bilingual associations of single words, that have been previously computed through statistical methods (see [Gaussier *et al.*, 1992]). Let $worden_1$ and $worden_2$ denote the two words composing an English candidate and, similarly, $wordfr_1$ and $wordfr_2$ those for a given French candidate. Each time a source and target candidate occur in aligned sentences, we define the association score between the two candidates as the sum of all the possible single-word-associations between the words composing the candidates:

$$\sum_{i,j} Assoc(worden_i, wordfr_j)$$

This method relies on the assumption that if two terms are translation of one another, then $wordfr_1$ is likely to be the translation of, or at least associated to, either $worden_1$ or $worden_2$, and so is $wordfr_2$. For example, both French words *station* and *terrienne* belong to the lists of single-word translation candidates associated to the English words *earth* and *station*.

4.3 Evaluation and results

To evaluate the results, we manually constructed a reference list which contains 1,238 correspondences of MWU2 we believe to be terms (appearing at least twice). For any list of candidate pairs to be tested (extracted and sorted according to some scores), we define its *precision* as the ratio of the number of elements found in the reference list to the total number of elements, and its *recall* as the ratio of the number of elements found in the reference list to the total number of elements in the reference list. In so far as these parameters are affected by the length of the list, we divided each list into units of the same length, considering the first 100 elements, the first 200 elements and so on, till we reach the end of one list. For each unit, we calculated its recall and its precision. The scores can then be compared in graphs showing the evolution of precision and recall with the number of candidates considered. Using such graphs, we found out that adding pattern probabilities gives only slightly better or equal results than the simple frequency method, which does not entirely satisfy our expectations as to their role on the quality of term alignment.

The second method presented above, which does not make use of pattern probabilities, behaves better for candidates of rank 1,000 and more, but worse for the first 100 best candidates. The examination of the top 100 candidates for both methods brings another surprise: only 6 term pairs are common to the two methods. Of course this could be due to the restriction to the very best candidates; but, suspecting that there is more to it, we combined the candidates from both methods and thus obtained an improvement of precision in all the range of candidates. The comparison between this combination of methods and the method involving pattern affinities is shown in figure 2. The plain line in figure 2 shows precision and recall for the first method involving pattern affinities (which is referred to as *normal* on the figure). We think that such results could be even better if we chose the right combination of two -or, possibly, more than two- scoring methods.

All in all, our method provides 70% precision for a list of more than 1,000 candidates, or 80% if one restricts to the topmost 500. It gives a good starting point for terminologists

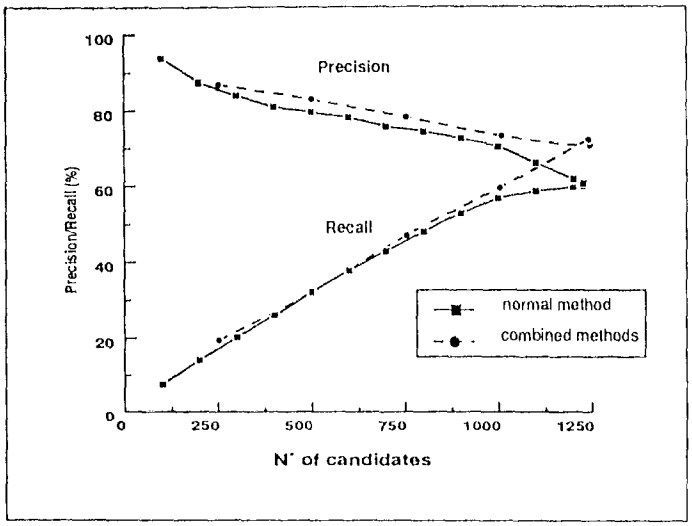

Figure 2: Precision and Recall when using first method vs. combined methods

seeking to extract bilingual terms from texts. It has to be noted that the remaining "wrong" alignments are generally somehow relevant, but either incomplete because one of the candidates has a length of more than 2, or trivial because we also extract pairs such as *following formula / formule suivante.*

5 Conclusion

For monolingual terminology extraction, we have first defined the linguistic specifications of multi-word unit terms. These specifications are used to extract candidate pairs from the corpora. From these pairs and using frequency counts, we have applied a range of statistical scores, in order to find the score that would best discriminate between terms and non-terms among the candidates. The simple frequency count of the pair turns out to be the best score. We then carried out bilingual terminology extraction using results from the monolingual step. Two scoring methods augmented by the use of prior knowledge about the bilingual correspondences of linguistic patterns were applied to bilingual pairs of monolingual candidate pairs. In this case, the best way to proceed is derived from the combination of the two methods.

With regard to the future, we will focus on monolingual extraction of MWU of length 3 and more (based on the identified base MWU and linguistic specifications on term composition), and on the improvement of term alignment in order to deal with terms of heterogeneous length.

References

[Benveniste, 1966] Benveniste É. (1966). Formes nouvelles de la composition nominale. *In Gallimard Ed., Problèmes de linguistique générale. pp. 163-173.*

[Bourigault, 1992] Bourigault D. (1992). Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. *Proceedings of the 14th International Conference on Computational Linguistics.*

[Brown *et al.*, 1988] Brown, P. F., Cocke, J., DellaPietra, S. A., DellaPietra, V. J., Jelinek, F., Mercer, R. L., and Roossin, P. S. (1988). A statistical approach to language translation. *Proceedings of the 12th International Conference on Computational Linguistics.*

[Church and Hanks, 1990] Church K. W. and Hanks P., 1990] (1990) Word Association Norms, Mutual Information, And Lexicography. *Computational Linguistics, Vol. 16, Number 1.*

[Clifford and Stephenson, 1975] Clifford H.T. and Stephenson W. (1975) *An Intro-*

duction to Numerical Classification, Academic Press.

[Daille, 1994] Daille B. (1994). Study and Implementation of combined techniques for Automatic Extraction of Terminology. *In The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Workshop at the 32nd Annual Meeting of the ACL.*

[Dunning, 1993] Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics, Vol. 19, Number 1.*

[Gaussier *et al.*, 1992] Gaussier E., Langé J.-M. and Meunier F. (1992). Towards Bilingual Terminology. *Proceedings of ALLC/ACH conference.*

[Hatcher, 1960] Hatcher A. J. (1960). An introduction to the analysis of English noun compounds. *In Word, 16, 356-373.*

[Jacquemin, 1991] Jacquemin C. (1991). Transformation des noms composés. *Thèse de doctorat en Informatique Fondamentale, Université Paris 7.*

[Mathieu-Colas, 1988] Mathieu-Colas M. (1988). Typologie des noms composés. *Technical Report of the "Programme de Recherches Coordonnées Informatique Linguistique", C.N.R.S.*

[Nkwenti-Azeh, 1992] Nkwenti-Azeh B. (1992). Positional and Combinational Characteristics of Satellite Communications Terms. *Interim Report, CCL-UMIST.*

[Smadja and McKeown, 1990] Smadja F. A. and McKeown K. R. (1990). Automatically Extracting And Representing Collocations For Language Generation. *Proceedings of the 28th annual Meeting of the ACL.*

[Van der Eijk, 1993] Van der Eijk P. (1993) Automating the Acquisition of Bilingual Terminology. *Proceedings of European ACL.*