Extracting Phraseology for Content Analysis and Document Retrieval

Thierry FONTENELLE, Hobscheid, Luxembourg

Abstract

This paper[1] describes a program which identifies the topic of a text by extracting the most relevant key words and phraseological sequences. The various factors taken into account to generate this list of terms and expressions are described (frequency of occurrence, classification as a function of the number of elements, processing of abbreviations, use of customisable stop lists…). The output can then be used by powerful search engines to retrieve topic-related texts which are believed to display a high degree of repetitivity, an essential criterion for building translation memory databases.

1 Introduction

The growth of the Internet and the availability of information in unprecedented quantities account for most of the recent efforts to develop efficient methods for retrieving information and sorting the wheat from the chaff. Locating a given document or a text dealing with a given topic on the basis of one or two keywords is often akin to looking for a needle in a haystack, especially when masses of data have to be browsed and examined.

Information Retrieval (IR) no longer needs to be explained because nearly everyone now has been confronted to the infuriating and often vain search for a document or a web site one knows should provide an answer to a particular problem. IR researchers regularly come up with new suggestions and methods to enable users to refine their requests and many search engines now incorporate linguistic knowledge which make it possible to retrieve documents even if they do not exactly match the users' requests (Grefenstette 1998). Applying morphological analysis to reduce a text to non-inflected lemma forms is probably the most widely used technique currently in use in IR. Resorting to synonym dictionaries and to thesauri in order to cast the net wider and locate documents which do not include a verbatim reference to the concept expressed in the query is also considered to provide better results. The main problem is that the linguistic resources which are required to improve IR (morphological analyzers, lexicons of related terms, concept hierarchies, etc) are usually not available when one starts dealing with texts written in languages other than English (although the situation might change with the advent of large semantic lexicons such as the (Euro)WordNet families (Vossen 1998)).

In this paper, we intend to show how simple techniques can be developed in order to identify the content of a given text and retrieve related texts, an important task in translation services in which the use of translation memories is becoming the rule rather than the exception. We will show in particular how significant terms and abbreviations can be extracted with a view to bringing out the contents of the document and identifying related texts which share the key terms with the source document.

2 Translation memory technology

The purpose of this section is not to describe the details of translation memory technology. Suffice it to say here that the basic principle which underlies the use of this fairly recent technology is to reuse what has already been translated in order to avoid duplicating efforts. In a nutshell, a translation memory system is a system which dynamically builds a database that remembers what has been translated and stores source and target sentences in order to allow the translator not to have to translate again a passage, a paragraph, a sentence or part of a sentence which has already been translated and for which aligned segments exist in the database (the 'memory') (see also Brockmann (1999) and Blatt (1998b) for more details about TM technology). The most important feature of TM systems is their ability to recognize segments which are not identical to what has to be translated, but which are nevertheless similar enough to ease the translator's task by providing him with a proposal which needs to be edited, updated or corrected. For instance, if the translator is to translate the following sentence, *Any animal suspected to be infected by a transmittible disease such as BSE may be slaughtered*, he may be interested in knowing that the database contains a very similar, though not identical sentence together with its translation, which may have been produced weeks or months before (e.g. *Animals suspected to be infected by a transmittible disease may be slaughtered*, with a translation into French such as*Les animaux soupçonnés d'avoir été infectés par une maladie transmissible peuvent être abattus*). In this case, fuzzy matching, a technique for finding data that has only a degree of similarity to the search argument (Blatt 1998b:93-98; Brockmann 1999:9), makes it possible to retrieve the already-translated English sentence from the memory, even though the source sentence which must be translated is used in the singular (*Any animal* vs. *Animals*) or material (*"such as BSE"*) has been added.

Translation memory technology rests upon the availability of parallel texts (Text B being the translated version of Text A). Special sofware to align these texts, i.e. to establish correspondences between source and target segments, usually at sentence level, is now available and comes in handy if the reference texts one wishes to use have not been translated with a TM system. The major problem is then to locate texts which are likely to contain a suitable number of 'hits', i.e. texts that are semantically close to the text that has to be translated and which can then be used to build a memory (assuming that texts which deal with the same type of topic are likely to be repetitive). In an environment such as the European Commission Translation Service, this task is akin to looking for a needle in a haystack since the textual archive which is put at the disposal of translators houses as many as 600,000 texts which can be searched with a powerful search and retrieval engine (see Scottini & Debart 1998 for more details on the use of this SDT*Vista* archive). Keywords and the traditional Boolean operators (and, or, not, proximity operators) can be used to nearly instantaneously retrieve a text which matches a given criterion (e.g. a given (set of) keyword(s) in an English text translated in 1999 for a given directorate-general).

3 ISABOUT: **Automatic content identification**

ISABOUT: is a prototype program designed to identify the most relevant and significant terms and keywords and key expressions which can in turn be used to optimize the search for a text dealing with the same topic as the text to be translated. The principle is fairly straightforward and is not very different from an earlier proposal made by Choueka (1988). The basic idea is to compute frequencies of occurrence of 2, 3 or 4-word expressions. The program uses as little linguistic information as possible so as to allow extension to as many languages as possible without requiring too much lexicon development. Another basic principle is that it also considers abbreviations and distinguishes them from isolated words by giving them more weight. It is indeed common knowledge that abbreviations are pervasive in specialized texts and since these abbreviations usually stand for complex terms, their identification is crucial when computing the topic of a text and looking for relevant keywords. The following sections describe the main features of the ISABOUT: program and a sample output of the program applied to a 20,000-word text dealing with bovine spongiform encephalopathy is given as appendix I.

3.1 Identification of abbreviations

Special routines were developed to extract abbreviations, taking into account the following factors:

- on the surface, abbreviations are single word items like other “isolated” words but they tell much more about the topic of a text than words taken in isolation (*ESB* [= encéphalopathie spongiforme bovine] and *MCJ* [= maladie de Creutzfeld-Jacob] appear 55 and 17 times, but are more relevant than *produits*, *consommateurs* or *cas*, which appear 57, 41 and 39 times respectively);

- abbreviations are usually capitalized;

- capitalized items preceded or followed by other capitalized words are ignored (because the risk of retrieving so-called “abbreviations” which are in fact capitalized chapter, paragraph or section headings is too high);

- some normalization is required in order to cater for variations in the presence or absence of periods which sometimes occur inside a text.

The latter point is essential to make sure superficial variations are ignored and the resulting figures are accurate. Consider the following KWIC lines excerpted from the text used in our experiment:

```
              un communiqué de l’     OMS     du 3 avril 1996
        figurant dans le communiqué     OMS     du 3 avril 1996
l’Organisation Mondiale de la Santé  (O.M.S.)  a organisé une
```

In the final computation, *OMS* (World Health Organisation) appears 3 times, irrespective of whether periods are used inside the abbreviation or not. If this routine was not applied, the frequency figures would not be so high and *OMS* would appear with a frequency of 2 only, which would send it further down the list of significant abbreviations.

3.2 Recurrent 2-, 3- and 4-word units are listed as a function of their frequency

In order to make sure that a 3-word term appearing at the top of the 3-word category is given more weight than a 2-word unit which appears at the bottom of its own category, irrespective of the absolute figures of frequency, the units are ranked *within* each category. As can be seen from Appendix I, an expression such as *étiquetage de la viande,* which appears 4 times in the category of 4-word expressions will probably be more relevant for our purpose than, say, *états membres* or *Union Européenne*, which appear 27 and 10 times respectively.

3.3 Use of a stop list

The program hardly makes use of linguistic knowledge. In order to avoid too much noise, however, this knowledge-poor technique can be enhanced by resorting to a language-specific stop list, containing the words one does not wish to take into account. In the present case, IS-ABOUT: ignores all terms and expressions starting or ending with determiners, auxiliaries, figures, prepositions, etc. This ensures that *de la viande* will not be considered, even if this three-word sequence appears a sizeable number of times, while *farine de viande*, appearing 5 times, is kept because the preposition is found inside the expression, which is allowed by the program.

Needless to say that the user is allowed to customize the stop list, for instance by adding words from open classes. The noun *Commission* occurs so frequently in texts produced by the European Commission that it is worth considering it as a stop word, since it is unlikely to play a major role in helping identify the topic of the text automatically. In another environment, for instance a Nato translation service, this list might include items such as *Nato* or *Shape*, which appear so often in texts produced by this organisation that they are no longer relevant in the present perspective.

3.4 Longest-match principle

The longest-match principle makes it possible to keep only longer strings if a shorter string is included in the longer one. In our text, *comité scientifique* is not kept as a potentially relevant keyword because it is included in *comité scientifique vétérinaire* and in *comité scientifique multidisciplinaire*, which are considered as relevant and appear at the top of the category of 3-word groups.

4 Automated topic identification vs. keyword identification

Many research projects aim at identifying the topic of a text automatically because this piece of information is often crucial in a word sense disambiguation perspective. Machine translation systems usually need this information in order to select the senses which are assigned a given subject field code. Lexical entries in MT systems are frequently tagged in terms of topical information, resorting to codes such as, say, "Food", "Sport", "Electronics", "Medicine", "Religion", etc. It has been shown that the subject code labels provided by general-language commercial dictionaries available in machine-readable form can be used to compute the topic of a text, which boils down to identifying the statistically most significant codes used in a text, without attempting to perform any kind of word sense disambiguation (see Walker & Amsler 1986 who use the LDOCE subject codes, and Jansen 1989 who computes the topic on the basis of subject labels from a bilingual dictionary). The general topic which is thus identified can subsequently be used to give priority to the word senses which bear the corresponding code.

The perspective adopted in this paper is different, however. While the technique described in the preceding paragraph would probably show that the text chosen for our experiment deals with something which would be called "Veterinary Medicine" or "Disease", or perhaps "Agriculture", this type of conclusion would be practically useless if this were to be used to identify other texts dealing with the same topic. Since we are primarily interested in identifying texts which display a high degree of repetitivity and which are semantically very close to other texts which have already been translated, we need to be much more precise than these general domain categories. The experiment showed that keeping, say, the first 5 to 8 lines of each sub-category (abbreviations, 4-word expressions, 3-word expressions. . .) produces a list of keywords and collocational expressions which can directly be submitted to a powerful search engine such as described by Scottini & Debart (1998), which makes it possible to zero in on lexically related texts featuring the same type of repetitivity and of lexical structures. It is just this sort of texts we are interested in when building a translation memory.

5 By-products

The algorithm described in this paper is based on a modular approach which makes it possible to reuse some of the components in other contexts. The routine which identifies abbreviations, for instance, is based on the use of regular expressions and can be used to extract the list of all the abbreviations used in a given text. The resulting list can in turn be used to automatically identify "gaps" in a terminological database, i.e. new abbreviations which do not appear in the database used and compiled by the translators and terminologists in a given domain. The full forms of abbreviations can also be discovered very easily, starting from the observation that the first occurrence of an abbreviation tends to be found between parentheses, to the immediate right of the complex term in its full, developed form (see Appendix II). The relevance of this information for a translator or a terminologist building a term base need not be demonstrated.

6 Conclusion

In this paper, we have described a program which identifies the most relevant keywords and key expressions, assuming that frequency of occurrence of recurring words, terms and phraseological expressions in a text can be used as clues to identify the topic of a text automatically. It is claimed that the list of keywords generated by this program can then be used to retrieve semantically related texts which can in turn be used to feed a translation memory and hence minimize the time spent on translating repetitive passages in specialized texts.

The technique presented here bears some resemblance with emerging technologies which enable translators and terminologists to build terminological databases by identifying and retrieving term candidates (Bourigault *et al.* 1996, Gaussier *et al.*1997). The latter programs normally do not make use of frequency data since the aim is not to sort the resulting “terms” as a function of their number of occurrences or the internal structure, giving greater weight to recurring longer items (which are considered more relevant for content identification than shorter items), as is the case here. Moreover, the technique presented here has been successfully applied to 4 languages (English, French, Dutch and German) and can easily be ported to other languages since it uses only very limited linguistic information and no part-of-speech tagger is necessary, although some lemmatization would probably enhance the retrieval of significant key words.

Notes

[1] This paper reports on work carried out while the author was working at the European Commission Translation Service. This institution disclaims any responsibility for the contents of this article and the views expressed here are not necessarily shared by the author’s former employer.

References

Blatt, A. (1998a): Translation Technology at the European Commission: Description of a workflow, *Terminologie & Traduction*, Luxembourg, 1.1998, pp.38-43.

Blatt, A. (1998b): Euramis Alignment and Translation Memory Technology, *Terminologie & Traduction*, Luxembourg, 1.1998, pp.74-101.

Bourigault, D., Gonzalez-Mullier, I. & Gros, C. (1996): LEXTER, A Natural Language Processing Tool for Terminology Extraction, in *Euralex’96 Proceedings*, Gothenburg, pp. 771-780.

Brockmann, D. (1999): Translation Memories as True Databases - Present and Future, *The ELRA Newsletter*, Vol.4, *N^o* 3, July-September 1999, pp.9-11.

Choueka, Yaacov (1988): Looking for needles in a haystack or locating interesting collocational expressions in large textual databases, *RIAO’88 Proceedings*, Cambridge, Mass., March 21-24, pp. 609-623.

Gaussier, E., Grefenstette, G. & Schulze, B.M. (1997): Traitement du langage naturel et recherche d’information: quelques exp expériences sur le français, *in Actes de FRANCIL’97*, Avignon, pp. 9-14.

Grefenstette, G. (ed.) (1998): *Cross-Language Information Retrieval*, Kluwer Academic Publishers, Dordrecht.

Jansen, J. (1989): Apport contrastif des dictionnaires généraux de la langue au problème de l’indexation automatique dans le discours techno-scientifique, *META*, 34/3, pp.412-427.

Scottini, M. & Debart, F. (1998)998): SDT*Vista*, la base documentaire en ligne du Service de traduction de la Commission, *Terminologie & Traduction*, Luxembourg, 1.1998, pp. 117-123.

Vossen, P. (ed.) (1998)*: EuroWordNet: A Multilingual Database with Lexical-Semantic Networks*, Kluwer Academic Publishers, Dordrecht.

Walker, D. & Amsler, R. (1986): The use of machine-readable dictionaries in sub-language analysis, in Grishman, R. & Kittredge, R. (eds) *Analyzing Language in Restricted Domains*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp.69-83.

Appendix I: Keywords and key terms from a text about BSE

Most relevant keywords and key terms appearing in a 20,000-word French text dealing with BSE (bovine spongiform encephalopathy, ESB in French). Five categories can be distinguished: abbreviations (deemed to be more relevant content-wise than anything else), 4-word expressions, 3-word expressions, 2-word expressions and single words, along a scale of decreasing relevance. Techniques to avoid noise are presented in sections 3.1 and 3.2 above, although 100 % precision can hardly be achieved. In the example below, it is clear that Weissmann is not an abbreviation, but the capitalization of proper names in this text led the program to erroneously consider it as an abbreviation.

55 esb
17 mcj
3 oms
3 v-mcj
2 weissmann
2 bse
2 vrm
2 csv
8 viande et d os
7 protection de la santé
6 produits de la filière
5 agent de l esb
5 esb au royaume uni
4 étiquetage de la viande
9 comité scientifique multidisciplinaire
9 comité scientifique vétérinaire
6 comité des consommateurs
5 système d étiquetage
5 alimentation des ruminants
5 encéphalopathie spongiforme bovine
5 farine de viande
5 humaine et animale
5 méthodes de production
4 provenant de mammifères
4 produits à base
4 spongiforme bovine esb
53 royaume uni
27 etats membres
12 mesures prises
11 santé publique
11 viande bovine
10 union européenne
8 chaîne alimentaire
8 moelle épinière
57 produits
41 consommateurs
39 cas
36 bovins
27 animaux
26 comité
23 scientifique
22 mars
22 conseil
21 recherche
21 scientifiques
20 pays
19 maladie
18 comité
17 vétérinaire
15 humaine
15 transmission
14 alimentation
14 avril
13 interdiction
12 communautaire
12 animale
12 juin
12 mois
12 niveau
12 agent

Appendix II

Kwic lines displaying abbreviations, appearing between parentheses at the end of the concordance, with their “decompacted” forms to their immediate left.

7 5 ENCEPHALOPATHIE SPONGIFORME BOVINE (ESB)
l’Organisation Mondiale de la Santé (O.M.S.)
qu’un lien entre l’encéphalopathie spongiforme bovine (ESB)
et la maladie de Creutzfeld Jacob (MCJ)
L’encéphalopathie spongiforme bovine (ESB)
La maladie de Creutzfeld Jacob (MCJ)
ont pris acte du vademecum sur l’encéphalopathie spongiforme bovine (ESB)
déclarent l’utilisation de viande reconstituée mécaniquement (VRM)