KNOWLEDGE ACQUISITION FROM A TEXT BY A LINGUISTIC AND STATISTICAL METHOD

Jean-David Sta
EDF, R&D
1, av du Général De Gaulle
92140 Clamart - France
E-mail : jean-david.sta@edf.fr

Yun-Chuang Chiao
DIAM–SIM Université de Paris 6
91, bd de l'Hôpital
75013 Paris - France
E-mail : ycc@biomath.jussieu.fr

Abstract
Faced with growing volume and accessibility of electronic textual information, information retrieval, and in general, automatic documentation require updated terminological resources that are ever more voluminous. A current problem is the automated construction of these resources from a corpus. Various linguistic and statistical methods to handle this problem are coming to light. One problem that has been less studied is that of updating these resources, in particular, of classifying a term extracted from a corpus in a subject field of an existing thesaurus. This experiment compares different models for representing a term from a corpus for its automatic classification in the subject fields of a thesaurus. Terms are first extracted from a corpus by combining a linguistic processing and statistical filtering The classification method used then is linear discriminatory analysis. The most effective model is the one based on mutual information between terms, which typifies the fact that two terms often appear together in the corpus, but rarely apart.

1 Introduction
In documentation, terminologies, thesauri and other terminological lists are reference systems which can be used for manual or automatic indexing. The quality of the result of the indexing process depends in large part on the quality of the terminology. Thus, applications downstream from the indexing depend on these terminological resources. The most thoroughly studied application is the information retrieval (IR). Here, the term provides a means for accessing information through its standardising effect on the query and on the text to be found. The term can also be a variable that is used in statistical classification or clustering processes of documents [BLO 92] or in selective dissemination of information [STA 93].

Textual information is becoming more and more accessible in electronic form. This accessibility is certainly one of the prerequisites for the massive use of natural language processing (NLP) techniques. These techniques applied on particular domains, often use terminological resources that supplement the lexical resources. The lexical resources (general language dictionaries) are fairly stable, whereas terminologies evolve dynamically with the fields they describe. Unfortunately, the abundance of electronic corpora and the relative maturity of natural language processing techniques have induced a shortage of updated terminological data. The various efforts in automatic acquisition of terminologies from a corpus stem from this observation, and try to know how can candidate terms be extracted from a corpus.

Another important question is how to position a term in an existing thesaurus. The question studied in this experiment concerns the positioning or classification of a term in a subject field of a thesaurus. This problem is very difficult for a human being to resolve when he is not an expert in the field to which the term belongs and one can hope that an automated classification process would be of great help. After a description of the corpus and the thesaurus, automatic indexing and terminology extraction are described. This preparatory processing allows the document to be represented as a set of terms (candidate terms and key words). A classification method is then implemented to classify a subset of 1,000 terms in the 49 themes and 330 semantic fields that make up the thesaurus. The 1,000 terms thus classified comprise the test sample that is used to evaluate three models for representing terms.

2 Data Preparation

2.1 Description of the Corpus

The corpus studied is a set of 10,000 scientific and technical documents in French (4,150,000 words). Each document consists of one or two pages of text. This corpus describes research carried out by the research division of EDF, the French electricity company. Many diverse subjects are dealt with: nuclear energy, thermal energy, home automation, sociology, artificial intelligence, etc. Each document describes the objectives and stages of a research project on a particular subject. Thus, the vocabulary used is either very technical, with subject field terms and candidate terms, or very general, with stylistic expressions, etc.

2.2 Description of the Thesaurus

The EDF thesaurus consists of 20,000 terms (including 6,000 synonyms) that cover a wide variety of fields (statistics, nuclear power plants, information retrieval, etc.). This reference system was created manually from corporate documents, and was validated with the help of many experts. Currently, updates are handled by a group of documentalists who regularly examine and insert new terms. One of the sources of new terms is the corpora. This thesaurus is composed of 330 semantic (or subject) fields included in 49 themes such as mathematics, sociology, etc.

2.3 Document Indexing

As a first step, the set of documents in the corpus is indexed. This consists of producing two types of indexes: candidate terms and descriptors. The candidate terms are expressions that may become terms. Descriptors are terms from the EDF thesaurus that are automatically recognised in the documents.

Terminological Filtering

In this experiment, terminological filtering is used for each document to produce terms that do not belong to the thesaurus, but which nonetheless might be useful to describe the documents. Linguistic and statistical terminological filtering are used. The method chosen for this experiment combines an initial linguistic extraction with statistical filtering [STA 95]. Generally, it appears that the syntactical structure of a term in French language is the noun phrase. In the EDF thesaurus, the terms are distributed as follows:

syntactic structure	example	%
Noun Adjective	**érosion fluviale**	**25.1**
Noun Preposition Noun	**analyse de contenu**	**24.4**
Noun	**décentralisation**	**18.1**
Proper noun	Chinon	6.8
Noun Preposition Article Noun	assurance de la qualité	3.2
Noun Preposition Noun Adjective	unité de bande magnétique	2.8
Noun Participe	puissance absorbée	2.2
Noun Noun	accès mémoire	2.1

Figure 1 : Distribution of the syntactical structures of terms

NP <- ADJECTIVE NP
NP <- NP ADJECTIVE
NP <- NP à NP
NP <- NP de NP
NP <- NP en NP
NP <- NP pour NP
NP <- NP NP

Figure 2 : The seven syntactic patterns for terminology extraction

Thus, term extraction is initially syntactical. It consists of applying seven recursive syntactic patterns to the corpus [OGO 94]. Linguistic extraction, however, is not enough. In fact, many expressions with a noun phrase structure are not terms. This includes general expressions, stylistic effects, etc. Statistical methods

can thus be used, in a second step, to discriminate terms from non-terminological expressions. Three indicators are used here:
Frequency : the more often an expression is found in the corpus, the more likely it is to be a term.
Variance : the more the occurrences in a document of an expression are scattered, the more likely it is to be a term.
Local density [STA 95]: the closer together the documents are that contain the expression, the more likely it is to be a term.

During this experiment, the terminological extraction ultimately produced 3,000 new terms that did not belong to the thesaurus.

Controlled Indexing

A supplementary way of characterising a document's contents is by recognising controlled terms in the document that belong to a thesaurus. To do this, an NLP technique is used [BLO 92]. Each sentence is processed on three levels: morphologically, syntactically, and semantically. These steps use a grammar and a general language dictionary. The method consists of breaking down the text fragment being processed by a series of successive transformations that may be syntactical (nominalisation, de-coordination, etc.), semantic (e.g., nuclear and atomic), or pragmatic (the thesaurus' synonym relationships are scanned to transform a synonym by its main form). At the end of these transformations, the decomposed text is compared to the list of documented terms of the thesaurus in order to supply the descriptors. Controlled indexing of the corpus supplied 4,000 terms (of 20,000 in the thesaurus). Each document was indexed by 20 to 30 terms. These terms, like the candidate terms, are used in the representation models described below.

3. Term Subject Field Discrimination

The problem of discrimination can be described as follows: A random variable X is distributed in a p-dimensional space. x represents the observed values of variable X. The problem is to determine the distribution of X among q distributions (the classes), based on the observed values x. The method implemented here is *linear discriminatory analysis*. Using a sample that has already been classified, discriminatory analysis can construct classification functions which take into account the variables that describe the elements to be classified. The classification function for an element x consists of choosing the class that has the highest probability. This function minimises the risk of classification error [RAO 65].

4 Description of the Experiment

The purpose of this experiment is to determine the best way to classify candidate terms from a corpus in semantic fields. The general principle is, firstly, to represent the candidate terms to be classified, then, to classify them, and finally, to evaluate the quality of the classification. The classification method is based on learning process, which requires a set of previously-classified terms (the learning sample manually classified). The evaluation also requires a test sample, a set of previously-classified terms which have to be automatically classified. The evaluation then consists of comparing the results of the classification process to the previous manual classification of the test sample.

4.1 Learning and Test Samples

The thesaurus terms found in the corpus were separated into two sets: a subset of about 3,000 terms which composed the learning sample, and a subset of 1,000 terms which composed the test sample. All these terms had already been manually classified by theme and semantic field in the thesaurus. The evaluation criteria is the *rate of well classified terms* calculated among the 1,000 terms of the test sample. The rate of well classified terms is the number of well classified terms divided by the number of classified terms.

4.2 Term Representation Models

The representation of the terms to be classified is the main parameter that determines the quality of the classification. Three models were evaluated.
The Term/Document Model : The term/document model uses the transposition of the standard document/term matrix of the vector space model [SAL 88].

The Term/Term Models : The term/term model uses a matrix where each line represents a term to be classified, and each column represents a thesaurus term recognised in the corpus, or a candidate term extracted from the corpus. At the intersection of a line and a column, two indicators have been studied.
Co-occurrences matrix: The indicator is the co-occurrence between two terms. Co-occurrence reflects the fact that two terms are found together in documents.
Mutual information matrix: The indicator is the mutual information between two terms. Mutual information ([CHU 89] and [FEA 61]) reflects the fact that terms are often found together, but rarely alone.

5 Results and Discussion

The main results concern three term representation models and two classifications: the first in 49 themes, and the second in 330 semantic fields. The criterion chosen for the evaluation is the well classified rate.

Method	Themes classification	Semantic fields classification
Term Document model	42.9	27.3
Term term model with co-occurrence	31.5	19.8
Term term model with mutual information	**89.8**	**65.2**

Figure 3 : Rate of well classified terms

Without a doubt, the term/term model with mutual information has the best performance. A detailed examination of the results shows that there is a wide dispersion of the rate of well classified terms depending on the field (the 49 themes or the 330 semantic fields). The explanation is that the documents in the corpus are essentially thematic. Thus, the vocabulary for certain fields in the thesaurus is essentially concentrated in a few documents. Classification based on mutual information is then efficient. On the other hand, certain fields are transverse (e.g., computer science, etc.), and are found in many documents that have few points in common (and little common technical vocabulary). Terms in these fields are difficult to classify.

Another problem with the method is connected to the representativeness of the learning sample. Commonly, for a given field, a certain number of terms are available (for example 20,000 terms in the EDF thesaurus). It is more rare for all these terms to be found in the corpus under study (4,000 terms found in this experiment). Thus, if a class (a theme or semantic field) is not well represented in the corpus, the method is unable to classify candidate terms in this class because the learning sample for this class is not enough.

References

[BLO 92] Blosseville M.J., Hebrail G., Monteil M.G., Penot N., "Automatic Document Classification: Natural Language Processing, Statistical Data Analysis, and Expert System Techniques used together ", ACM-SIGIR'92 proceedings, 51-58,1992.
[CHU 89] Church K., "Word Association Norms, Mutual information, and Lexicography ", ACL 27 proceedings, Vancouver, 76-83, 1989.
[DAS 90] Dasarathy B.V., "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques", IEEE Computer Society Press, 1990.
[FEA 61] Fano R., "Transmission of Information", MIT Press, Cambridge, Massachusetts, 1961.
[OGO 94] Ogonowski A., Herviou M.L., Monteil M.G., "Tools for extracting and structuring knowledge from text", Coling'94 proceedings, 1049-1053, 1994.
[RAO 65] Rao C.R., "Linear Statistical Inference and its applications ", 2nd edition, Wiley, 1965.
[SAL 88] Salton G., « Automatic Text Processing : the Transformation, Analysis, and Retrieval of Information by Computer », Addison-Wesley, 1988.
[STA 93] Sta J.D., "Information filtering : a tool for communication between researches", INTERCHI'93 proceedings, Amsterdam, 177-178, 1993.
[STA 95] Sta J.D., "Comportement statistique des termes et acquisition terminologique à partir de corpus", T.A.L., Vol. 36, Num. 1-2, 119-132, 1995.