Knowledge Acquisition:
Classification of Terms in a thesaurus from a Corpus

STA Jean-David

EDF, Direction des Etudes et des Recherches
1, av. de Général de Gaulle
92140 Clamart - France

(33) 1 47.65.49.58   E-mail : jd.sta@der.edfgdf.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 (e.g., terminologies, thesauri, glossaries, etc.) 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, discipline or branch of an existing thesaurus. This is the first step in positioning a term extracted from a corpus in the structure of a thesaurus (generic relations, synonymy relations …). This is an important problem in certain disciplines in which knowledge, and, in particular, vocabulary is not very stable over time, especially because of neologisms.

This experiment compares different models for representing a term from a corpus for its automatic classification in the subject fields of a thesaurus. The classification method used is linear discriminatory analysis, based on a learning sample. The models evaluated here are: a term/document model where each term is a vector in document vector space, two term/term models where each term is a vector in term space, and the coordinates are either the co-occurrence, or the mutual information between terms. 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.

I. Introduction

In documentation, terminologies, thesauri and other terminological lists are reference systems which can be used for manual or automatic indexing. Indexing consists of recognising the terms in a text that belong to a reference system; this is called controlled indexing. The quality of the result of the indexing process depends in large part on the quality of the terminology (completeness, consistence …). 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] and [STA 95a]), or in selective dissemination of information, in which it is used to bring together a document to be disseminated and its target [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. In particular, the disciplines of information processing (computers, etc.) and biology or genetics are characterised today by an extraordinary terminological activity.

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 answer the following question: "How can candidate terms be extracted from a corpus?"

Another important question is how to position a term in an existing thesaurus. That question can itself be subdivided into several questions that concern the role of the standard relationships in a thesaurus: synonymy, hyperonymy, etc. The question studied in this experiment concerns the positioning or classification of a term in a subject field or semantic field of a thesaurus. This is the first step in a precise positioning using the standard relationships 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.

To classify a term in a subject field can be considered similar to word sense disambiguation (WSD) which consists in classifying a word in a conceptual class (one of its senses). The difference is that, in a corpus, a term is generally monosemous and a word is polysemous. Word sense disambiguation uses a single context (generally a window of a few words around the word to be disambiguated) as input to predict its sense among a few possible senses (generally less than ten). Term subject field discrimination uses a representation of the term calculated on the whole corpus in order to classify it into about 330 subject fields in this experiment.

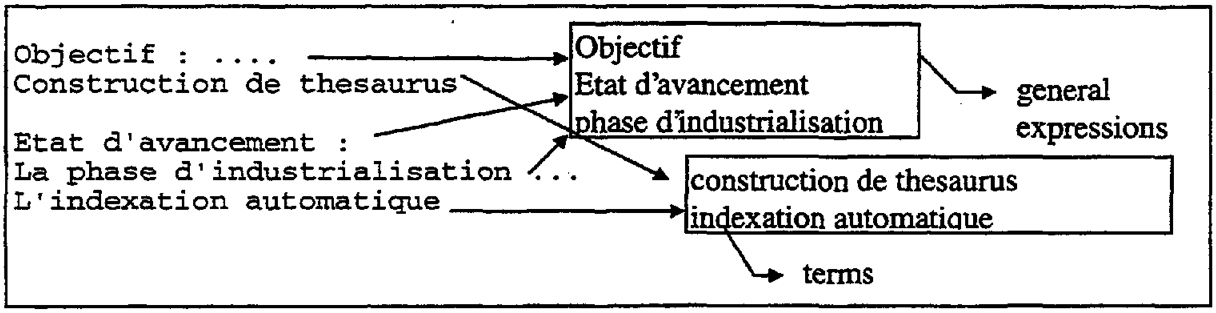
The experiment described here was used to evaluate different methods for classifying terms from a corpus in the subject fields of a thesaurus. After a brief description of the corpus and the thesaurus, automatic indexing and terminology extraction are described. Linguistic and statistical techniques are used to extract a candidate term from a corpus or to recognise a term in a document. 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.

II. Data Preparation

II.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. These documents are used to plan EDF research activity.

Thus, the vocabulary used is either very technical, with subject field terms and candidate terms, or very general, with stylistic expressions, etc.

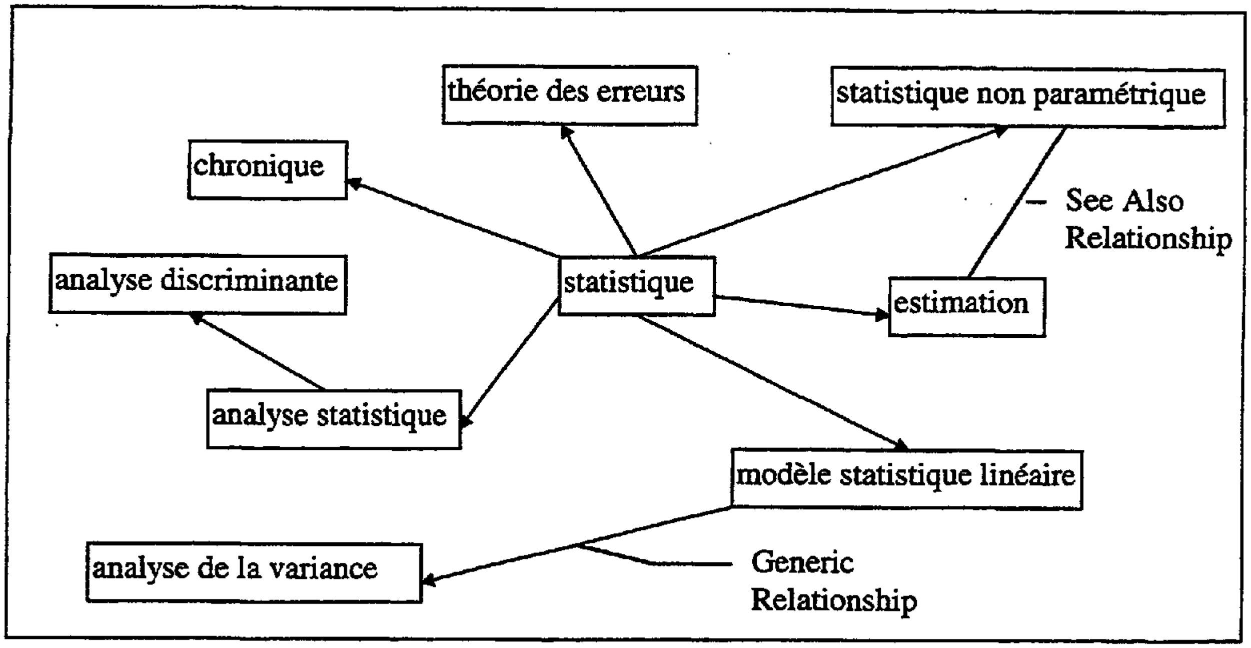


A document with terms and general expressions

II.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. A linguistic and statistical extractor proposes candidate terms for validation by the documentalists. After validation, the documentalists must position the selected terms in the thesaurus. It's a difficult exercise because of the wide variety of fields.

The thesaurus is composed of 330 semantic (or subject) fields included in 49 themes such as mathematics, sociology, etc.



Extract from the "statistics" semantic field from the EDF thesaurus

This example gives an overview of the various relations between terms. Each term belongs to a single semantic field. Each term is linked to other terms through a generic relation (arrow) or a neighbourhood relation (line). Other relations (e.g., synonym, translated by, etc.) exist, but are not shown in this example.

II.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, and are submitted to an expert for validation. Descriptors are terms from the EDF thesaurus that are automatically recognised in the documents.

II.3.1. 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. Moreover, these expressions are candidate terms that are submitted to experts or documentalists for validation.

Linguistic and statistical terminological filtering are used. The method chosen for this experiment combines an initial linguistic extraction with statistical filtering [STA 95b].

Linguistic Extraction

Generally, it appears that the syntactical structure of a term in French language is the noun phrase. For example, in the EDF thesaurus, the syntactic structures of 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

Distribution of the syntactic structures of terms

Thus, term extraction is initially syntactical. It consists of applying seven recursive syntactic patterns to the corpus [OGO 94].

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

The seven syntactic patterns for terminology extraction

Statistical Filtering

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: This is based on the fact that the more often an expression is found in the corpus, the more likely it is to be a term. This statement must be kept in proportion, however. Indeed, it seems that a small number of words (usually, very general uniterms) are very frequent, but are not terms.

- Variance: This is based on the idea that the more the occurrences in a document of an expression are scattered, the more likely it is to be a term. This is the most effective indicator. Its drawback is that it also highlights large noun phrases in which the terms are included.

- Local density [STA 95b]: This is based on the idea that the closer together the documents are that contain the expression, the more likely it is to be a term. The local density of an expression is the mean of the cosines between documents which contain the given expression. A document is a vector in the Document Vector Space where a dimension is a term. This indicator highlights a certain number of terms that are not transverse to the corpus, but rather concentrated in documents that are close to each other. Nonetheless, this is not a very effective indicator for terms that are transverse to the corpus. For example, terms from computer science, which are found in a lot of documents, are not highlighted by this indicator.

Results of the Terminological Extraction

During this experiment, the terminological extraction ultimately produced 3,000 new terms that did not belong to the thesaurus. These new terms are used in the various representation models described below. The initial linguistic extracting produced about 50,000 expressions.

II.3.2. 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.

Results of the Controlled Indexing

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 documented terms, like the candidate terms, are used in the representation models described below. The quality of the indexing process is estimated at 70 percents (number of right terms divided by number of terms). The wrong terms are essentially due to problems of polysemy. Indeed some terms (generally uniterms) have multiple senses (for example "BASE") and produce a great part of the noise.

III. Term Subject Field Discrimination

III.1. Word Sense Disambiguation and Term Subject Field Discrimination

The discrimination of word senses is a well known problem in computational linguistics [YNG 55]. The problem of WSD is to build indicators describing the different senses of a word. Given a context of a word, these indicators are used to predict its sense. Face to the difficulty of manually building these indicators, researchers have turned to resources such as machine-readable dictionaries [VER 90] and corpora [YAR 92].

WSD and term subject field discrimination from corpora can be considered similar in the way that they are both a problem of classification into a class (a sense for a word and a subject field for a term). Nevertheless, the problem is statistically different. In one case, a word is represented by a few variables (its context) and is classified into one class chosen among a few classes. In the other case, a term is represented by hundred of variables (one of the models described in chapter IV) and is classified into a class chosen among hundred of classes.

III.2. Linear Discriminatory Analysis

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. Each element x to be classified is described by a binary vector x= (x1, x2, ..., xi, ..., xp) where xi=1 or xi=0.

xi=1 means the variable xi describes the term x.
xi=0 means the varaible xi does not describe the term x.

The probability that an element x is in a class c is written:

$P(C=c \mid X=x)$ where C is a random variable and X is a random vector.

Using Bayes formula, it may be deduced that:

$$P(C=c \mid X=x) = (P(C = c) . P(X = x \mid C = c)) / P(X = x)$$

There are three probabilities to estimate:

- Estimate P(C = c)

$P(C = c)$ is estimated by nc / n where:

nc is the number of elements of the class c
n is the number of elements in the sample

- Estimate $P(X = x)$

This estimate is simplified by normalising the probabilities to 1.

- Estimate $P(X = x \mid C = c)$

For this estimate, we assume that the random variables $X1, X2, \ldots Xm$ are independent for a given class c. This leads to:

$P(X = x \mid C = c) = \Pi P(Xi = xi \mid C = c)$ and

$P(Xi = 1 \mid C = c)$ is estimated by $nc,i / nc$
$P(Xi = 0 \mid C = c)$ is estimated by $1 - nc,i / nc$

where nc,i is the number of elements in the sample which are in class c, and for which $xi =1$, and nc is the number of elements of the sample in class c.

Once all the probabilities are estimated, 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].

IV. 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.

IV.1. Learning and Test Sample

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.

Rate of Well Classified Terms

The evaluation criteria is the rate of well classified terms calculated among the 1,000 terms of the test sample.

Rate of well classified terms = number of well classified terms divided by the number of classified terms.

IV.2. Term Representation Models

The representation of the terms to be classified is the main parameter that determines the quality of the classification. Indeed, this experiment showed that, for a single representation model, there is no significant difference between the results of the various classification methods. By example, the nearest neighbours method (KNN) [DAS 90] was tested without any significant difference. The only parameter that truly influences the result is the way of representing the terms to be classified. Three models were evaluated. The first is based on a term/document approach, and the two others by a term/term approach.

IV2.1. The Term/Document Model

The term/document model uses the transposition of the standard document/term matrix. Each line represents a term, and each column a document. At the intersection of a term and a document, there is 0 if the term is not in the document in question, and 1 if it is present.

The standard document/term matrix showed its worth in the Salton vector model [SAL 88]. It can therefore be hoped that the documents that contain a term provide a good representation of this term for its classification in a field.

IV.2.2. 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 two terms are often found together in documents, but rarely alone. MI(x,y) is the mutual information between terms x and y, and is written:

MI(x,y) = log2(P(xy) /P(x) . P(y))

where P(x,y) the probability of observing x and y together and P(x) the probability of observing x, P(y) the probability of observing y.

In the two cases, the matrix has to be transformed into a binary matrix. The solution is to choice a threshold under which the value is put to 0 and above which the value is put to 1. Lots of values had been tested. The best classification for the co-occurrence matrix is obtained for a threshold of three. The best classification for the mutual information matrix is obtained for a threshold of 0.05.

Results

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 classificati on
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**

Rate of Well Classified Terms

There is a significant difference between the term/term model with mutual information and the other two models. The good rates (89.8 and 65.2) can be improved if the system proposes more than one class. In the case of 3 proposed classes (sorted by descending probabilities), the probability that the right class is in these classes is estimated respectively by 97.1 and 91.2 for the themes and the semantic fields.

Discussion

Without a doubt, the term/term model with mutual information has the best performance. Nonetheless, these good results must be qualified.

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 320 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.

Through this experiment, an automatic classification of 300 candidate terms in 330 semantic fields was proposed to the group that validates new thesaurus terms. This classification was used by the documentalists to update the EDF thesaurus. Each term was proposed in three semantic fields (among 330) sorted from the highest probability (to be the right semantic field of the term) to the lowest.

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 95a] Sta J.D., "Document expansion applied to classification : weighting of additional terms", ACM-SIGIR'95 proceedings, Seattle, 177-178, 1995.

[STA 95b] 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.

[VER 90] Veronis J., Ide N., "Word Sens Disambiguation with Very Large Neural Networks Etracted from Machine Radable Dictionnaries" COLING'90 proceedings, 389-394, 1990.

[YAR 92] Yarowsky D., "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", COLING'92 proceedings, 454-460, 1992.

[YNG 55] Yngve V., "Syntax and the Problem of Multiple Meaning" in Machine Translation of Languages, William Lock and Donald Booth eds., Wiley, New York, 1955.