Terminological variation, a means of identifying research topics from texts

Fidelia IBEKWE-SANJUAN
CRISTAL-GRESEC, Stendhal University, Grenoble France
and
Dept. of Information & Communication
IUT du Havre - B.P. 4006 - 76610 Le Havre France
E-mail : fidelia@iut.univ-lehavre.fr

Abstract

After extracting terms from a corpus of titles and abstracts in English, syntactic variation relations are identified amongst them in order to detect research topics. Three types of syntactic variations were studied : permutation, expansion and substitution. These syntactic variations yield other relations of formal and conceptual nature. Basing on a distinction of the variation relations according to the grammatical function affected in a term - head or modifier - term variants are first clustered into connected components which are in turn clustered into classes. These classes relate two or more components through variations involving a change of head word, thus of topic. The graph obtained reveals the global organisation of research topics in the corpus. A clustering method has been built to compute such classes of research topics.

Introduction

The importance of terms in various natural language tasks such as automatic indexing, computer-aided translation, information retrieval and technology watch need no longer be proved. Terms are meaningful textual units used for naming concepts or objects in a given field. Past studies have focused on building term extraction tools : TERMINO (David S. & Plante P. 1991), LEXTER (Bourigault D. 1994), ACABIT (Daille 1994), FASTR (Jacquemin 1995), TERMS (Katz S.M. & Justeson T.S. 1995). Here, term extraction and the identification of syntactic variation relations are considered for topic detection. Variations are changes affecting the structure and the form of a term producing another textual unit close to the initial one e.g. *dna amplification* and *amplification fingerprinting of dna*. Variations can point to terminological evolution and thus to that of the underlying concept. Topic is used in its grammatical sense, i.e. the head word in a noun phrase. In the above term, *fingerprinting* is the topic (head word) and *dna amplification* its properties (modifiers). However, a topic cannot appear by chance in specialised litterature, so this grammatical definition needs to be backed up by empirical evidence such as recurrence of terms sharing the same head word. We constituted a test corpus of scientific abstracts and titles in English from the field of plant biotechnology making up ≅29000 words. These texts covered publications made over 13 years (1981-1993). We focused on three syntactic variation types occurring frequently amongst terms : permutation, substitution and expansion (§2). Tzoukermann E. Klavans J. and Jacquemin C. (1997) extracted morpho-syntactic term variants for NLP tasks such as automatic indexing. They accounted for a wide spectrum of variation producing phenomena like the morpho-syntactic variation involving derivation in *tree cutting* and *trees have been cut down*[1].

We focused for the moment on terms appearing as noun phrases (NP). Although term variants can appear as verb phrases (VP), we believe that NP variants reflect more terminological stability thus a real shift in topic (*root hair* → *root hair deformation*) than their VP counterpart (*root hair* → *the root hair appears deformed*). Also, our application - research topic identification - being quite sensitive, requires a careful selection of term variants types depending on their interpretability.

[1] Examples taken from Tzoukermann et al. (1997).

This is to avoid creating relations between terms which could mislead the end-user, typically a technological watcher, in his task. For instance how do we interpret the relation between *concept class* and *class concept* ? Also, our aim is not to extract syntactic variants *per se* but to identify them in order to establish meaningful relations between them.

1 Extracting terms from texts

1.1 Morpho-syntactic features

Term extraction is based on their morpho-syntactic features. The morphological composition of NP terms allows for a limited number of categories mostly nouns, adjectives and some prepositions. Terms can appear under two syntactic structures : compound (*the specific alfalfa nodulation*) or syntagmatic (*the specific nodulation of alfalfa*). Since terms are used for naming concepts and objects in a given knowledge field, they tend to be relatively short textual units usually between 2-4 words though terms of longer length occur (*endogeneous duck hepatitis B virus*). In this study, we fixed a word limit of 7 not considering determiners and prepositions.

Based on these three features, morphological make-up, syntactic structure and length, clauses are processed in order to extract complex terms rather than atomic ones. The motivation behind this approach is that complex terms reveal the association of concepts, hence they are more relevant for the application we are considering. A fine-grained term extraction strategy would isolate the concepts and thus lose the information given by their associations in the corpus. For this reason, we could not consider the use of an existing term extraction tool and thus had to carry out a manual simulation of the term extraction phase. NP splitting rules take into account the lexical nature of the constituent words and their raising properties (i.e. derived nouns as opposed to non-derived ones). Furthermore, following the empirical approach successfully implemented by Bourigault (1994), we split complex NPs only after a search has been performed in the corpus for occurrences of their sub-segments in unambiguous situations, i.e. when the sub-segments are not included in a larger segment. This favours the extraction of pre-conceived textual units possibly corresponding to domain terms. However morpho-syntactic features alone cannot verify the terminological status of the units extracted since they can also select non terms (see Smadja 1993). For instance *root nodulation* is a term in the plant biotechnology field whereas *book review* also found in the corpus is not. Thus in the first stage, the terms extracted are only plausible candidates which need to be filtered in order to eliminate the most unlikely ones. This filtering takes advantage of lexical information accessible at our level of analysis to fine-tune the statistical occurrence criterion which used alone, inevitably leads to a massive elimination.

1.2 Splitting complex noun phrases

An NP is deemed complex if its morpho-syntactic features do not conform to that specified for terms, e.g. *oxygen control of nitrogen fixation gene expression in bradyrhizobium japonicum* a title found in our corpus. Its corresponding syntactic context is : NP_1_of_NP_2_$prep_1$_NP_3 where NP is a recognised noun phrase, $prep_1$ refers to the class of preposition not containing *of* and often found in the morphological composition of terms (*for, by, in, from, with*). Normally, exploiting syntactic information on the raising properties of the head noun (*control*) and following the distributional approach, the above segment will be split thus :

$\rightarrow NP_1$
$\rightarrow NP_2$
$\rightarrow NP_3$

But this splitting is only performed if no sub-segment of the initial one occurred alone in the corpus. This search yielded *nitrogen fixation gene expression* and *bradyrhizobium japonicum* which both occurred more than 6 times in the corpus. Their existence confirms the relevance of our splitting rule which would have yielded the same result : *oxygen control; nitrogen fixation gene expression; bradyrhizobium japonicum*

Altogether, 4463 candidate terms were extracted from our corpus and subjected to a filtering process which combined lexical and statistical criteria. The lexical criterion consisted in eliminating terms that contained a determiner other than *the* that remained after the splitting phase. Only this determiner can occur in a term as it has the capacity, out of context, to refer to a concept or object in a knowledge field, i.e. the use

of the variant *the low-line* instead of the full term *low fertility droughtmaster line*[2]. The statistical criterion consisted in eliminating terms starting with *the* and appearing only once. These two criteria enabled us to eliminate 30% (1304) candidates and to retain 70% (3159) which we consider to be likely terminological units. We are aware that this filtering procedure remains approximate and cannot eliminate bad candidates like *book review* whose morphological and lexical make-up correspond to those of terms. But we also observe that such bad candidates are naturally filtered out in later stages as they rarely possess variants and thus will not appear as research topics (see §4).

2 Identifying syntactic variants

Given the two syntactic structures under which a term can appear - compound or syntagmatic - we first pre-processed the terms by transforming those in a syntagmatic structure into their compound version. This transformation is based on the following noun phrase formation rule for English :

$D\ A\ M_1\ h\ p\ m\ M_2 \rightarrow D\ A\ m\ M_2\ M_1\ h$

where *D*, *A* and *M* are respectively strings of determiner, adjective and words whose place can be empty, *h* is a head noun, *m* is a word and *p* is a preposition. Thus, the compound version of *the specific nodulation of alfalfa* will give *the specific alfalfa nodulation*. This transformation does not modify the original structure under which a term occurred in the corpus. It only serves to furnish input data to the syntactic variation identification programs. This transformation which is equivalent to permutation (§2.1) is the linguistic relation which once accounted for, reveals the formal nature of the other types of syntactic variations. Also, it enables us to detect variants in the two syntactic structures thus accounting for syntactic variants such as defined in Tzoukermann et al. (1997). In what follows, t_1 and t_2 are terms.

2.1 Permutation (Perm)

It marks the transformation of a term, from a syntagmatic structure to a compound one :

$t_1 = A\ N\ M_1\ h\ p\ m\ M_2$

$t_2 = A\ m\ M_2\ N\ M_1\ h$

where t_1 is really found in the corpus, *N* is a string of words that is either empty or a noun. 37 terms were concerned by this relation. Some examples are given in Table 1.

2.2 Substitution (Sub)

It marks the replacing of a component word in t_1 by another word in t_2 in terms of equal length. Only one word can be replaced and at the same position to ensure the interpretability of the relation. We distinguished between modifier and head substitution.

- *Modifier substitution (M-Sub) :*
  t_2 is a substitution of t_1 if and only if :
  $t_1 = M_1\ m\ M_2\ h$ and $t_2 = M_1\ m'\ M_2\ h$
  with $m' \neq m$
- *Head substitution (H-Sub) :*
  t_2 is a substitution of t_1 if and only if :
  $t_1 = M\ m\ h$ and $t_2 = M\ m\ h'$
  with $h' \neq h$

Tzoukermann et al. (1997) considered *chemical treatment against disease* and *disease treatment* as substitution variants whereas, in our study, after transformation, they would be a case of left-expansion (L-Exp). Examples of head and modifier substitutions are given in Table 2. 1543 terms shared substitution relations : 1084 in the modifier substitution and 872 in the head substitution. The same term can occur in both categories.

2.3 Expansion (Exp)

Expansion is the generic name designating three elementary operations of word adjunction in an existing term. Word adjunction can occur in three positions : left, right or within. Thus we have left expansion, right expansion and insertion respectively.

- *Left expansion (L-Exp) :*
  t_2 is a left-expansion of t_1 if and only if :
  $t_1 = M\ h$ and $t_2 = M'\ m'\ M\ h$
- *Right expansion (R-Exp) :*
  t_2 is a right-expansion of t_1 if and only if :
  $t_1 = M\ h$ and $t_2 = M\ h\ M'\ h'$
- *Insertion (Ins) :*
  t_2 is an insertion of t_1 if and only if :
  $t_1 = M_1\ m\ M_2\ h$
  $t_2 = M_1\ m\ m'\ M'\ M_2\ h$

[2] It apparently refers to a breed (line) of cattle.

Examples of each sub-type of expansion are given in Table 3.

Some terms combine the two types of expansion - left and right expansions (noted LR-Exp), for example *root of bragg* → *root exudate of soyabean cultivar bragg*. These complex expansion variants were also identified. A total of 1014 terms were involved in the expansion variation relations. Altogether, 82% (2593 out of 3159) terms were involved in the three types of syntactic variations studied showing the importance of the phenomena amongst terms.

Syntagmatic structure	Compound structure
accession of azolla-anabaena	*azolla-anabaena accession*
avirulent strain of pseudomonas syringae	*avirulent pseudomonas syringae strain*
curling of root hair	*root hair curling / root-hair curling*
excision of nodule	*nodule excision*
the specific nodulation of alfalfa	*the specific alfalfa nodulation*

Table 1. Examples of permutation variants identified in the corpus.

Head substitution variants	Modifier substitution variants
nodule development *regulation* nodule development *arrest* nodule development *consequence*	*alfalfa* root hair *curled* root hair *lucerne* root hair
infection thread *development* infection thread *formation* infection thread *initiation*	*characteristic* dna fingerprinting *conventional* dna fingerprinting *complex* dna fingerprinting
nodulation of soybean mutant *isolation* of soybean mutant *property* of soybean mutant	*enzymatic* amplification of dna amplification of *genomic* dna

Table 2. Some head and modifier substitution variants identified in the corpus.

Left expansion	Right expansion	Insertion
self-licking → *refractory* self-licking *stereotypic* self-licking	blue light → blue light-*induced expression* blue light *induction* blue light *induction experiment*	conserved domain → conserved *central* domain conserved *protein* domain
nitrogenase activity → nitrogenase activity of *cv. bragg* nitrogenase activity of *nitrate* nitrogenase activity of *nts382* nitrogenase activity of *soyabean*	immigrant of eastern countries → immigrant *children* of eastern countries[3]	fast staining of dna→ fast *silver* staining of dna

Table 3. Examples of expansions variants identified in the corpus.

The programs identifying syntactic variants were written in the Awk language and implemented on a Sun Sparc workstation.

Syntactic variations possess formal properties such as symmetry and antisymmetry. Permutation and substitution engender a *symmetrical relation* between terms, e.g. *genomic dna* σ *template dna*.

[3] This example is fictitious.

Expansion engenders an *antisymmetrical* or *order relation* between terms, for instance *nitrogen fixation<nitrogen fixation gene<nitrogen fixation gene activation*. These two formal properties will form the second level for differentiating variation relations during clustering (see §4).

3 Conceptual properties of syntactic variations

Syntactic variations yield conceptual relations which can reveal the association of concepts represented by the terms. We observed three conceptual relations : *class_of*, *equivalence*, *generic/specific*.

- *Class_of*

Substitution (Sub) engenders a relation between term variants which can be qualified as "class_of". Modifier substitution groups properties around the same concept class : *template dna*, *genomic dna*, *target dna* are properties associated to the class of concept named *"dna"*. Head substitution groups concepts or objects around a class of property : *dna fragment*, *dna sequence*, *dna fingerprinting* are concepts associated to the class of property named *dna*. This relation does not imply a hierarchy amongst terms thus somehow reflecting the symmetrical relation engendered on the formal level.

- *Equivalence*

Permutation engenders a conceptual equivalence between two variants which partially echoes the formal symmetry, e.g. *dna fragment* ≡ *fragment of dna*.

- *Generic / specific*

Expansion, all sub-types considered, engenders a generic/specific relation between terms which echoes the antisymmetrical relation observed on the formal level. Expansion thus introduces a hierarchy amongst terms and allows us to construct paradigms that may correspond to families of concepts or objects (R-Exp, LR-Exp) or families of properties (L-Exp, Ins). Jacquemin (1995) reported similar conceptual relations for insertion and coordination variants.

4 Identifying topics organisation

We built a novel clustering method - Classification by Preferential *Clustered* Link (CPCL) - to cluster terms into classes of research topics. First we distinguished two categories of variation relations : those affecting modifier words noted *COMP* (M-Sub, L-Exp, Ins) and those affecting the head word noted *CLAS* (H-Sub, LR-Exp, R-Exp).

The need to value the variation relations may arise if a type (symmetrical or antisymmetrical) is in the minority. To preserve the information it carries, a default value is fixed for this minority type. The value of the majority type is then calculated as its proportion with regard to the minority type. In our corpus, *Exp* (antisymmetrical) relations were in minority compared to *Sub* (symmetrical relations). Their default value was set at 1. The value of Sub relations was then given by the ratio *Exp/Sub* where *Exp* (respectively *Sub)* is the total number of expansions relations (respectively substitutions) between terms in the corpus. This valuing of variation relations highlights a type of information that would otherwise be drowned but is not a mandatory condition for the clustering algorithm to work.

COMP relations structure term variants around the same head word thus forming components representing the paradigms in the corpus. These paradigms typically correspond to isolated topics (see Table 4 hereafter). The strength of the link between two components P_i and P_j is given by the sum of the value of variation relations between them. More formally, we define the *COMP* relation between terms as : t_i *COMP* t_j iff t_i and t_j share the same head word and if one is the variant of the other. The transitive closure *COMP** of *COMP* partitions the whole set of terms into components. These components are not isolated and are linked by transversal *CLAS* relations implying a change of head word, thus bringing to light the associations between research topics in the corpus.

CLAS relations cluster components basing on the following principle : *two components P_i and P_j are clustered if the link between them is stronger than the link between either of them and any other component P_k which has not been clustered neither with P_i nor with P_j.* We call *classification*, a partition of terms in such classes. An efficient algorithm has been implemented in Ibekwe-SanJuan (1997) which seeks growing series of

such classifications. These series represent more or less fine-grained structurings of the corpus. A more formal description of the CPCL method can be found in Ibekwe-SanJuan (1998).
Table 4 shows a component and a class.

The component formed around the head word *hair* reveals the properties (modifiers) associated with this topic but does not tell us anything about its association other topics. The class on the other hand reveals the association of *hair* with other topics.

A component	A class of terms
alfalfa root hair	alfalfa root *hair*
curled root hair	concomitant root *hair* curling
deformed root hair	curled root *hair*
lucerne root hair	deformed root *hair*
root hair	*hair* deformation
	lucerne root *hair*
	occasional *hair* curling
	root deformation
	root *hair*
	root *hair* curling
	root *hair* deformation
	some root *hair* curling

Table 4. A component and a class.

The graph in Figure 1 hereafter shows the global organisation of classes obtained from the classification of the entire corpus (2593 syntactic term variants).
External links between classes are given by bold lines for R-Exp and LR-Exp, dotted lines portray head-substitution H-Sub. Only one term from each class is shown for legibility reasons. We observe that classes like 17, 19, 18 and 9 have a lot of external links and seem to be at the core of research topics in the corpus. Classes like 12, 3 and 13 share strong external links with a single class which could indicate privileged thematic relations. The unique link between class 3 and 19 is explained by the fact that 3 represented an emerging topic[4] at the time the corpus was constituted (1993) : the research done around a new *gene* type (the *klebsiella pneumoniae nifb gene*). So it was relevant that this class be strongly linked to class 19 without being central. Also, class 10 represented an emerging topic in 1993 : the research for *retrotransposable elements* which enables the passing from one *gene* to another. Research topics evolution and transformation can be traced through a chronological analysis of clustered term variants (see Ibekwe-SanJuan 1998). The results obtained can support scientific and technological watch activities.

Concluding remarks

Syntactic variation relations are promising linguistic phenomena for tracking topic evolution in texts. However, being that clustering is based on syntactic variation relations, the CPCL method cannot detect topics related through semantic or pragmatic relations. For instance, the topic depicted by class 8 (*glycine max*) should have been related to topic 20 (*lucerne plant*) from a semantic viewpoint. Their separation was caused by the absence of syntactic variations between the constituent terms. Such relations can be brought to light only if further knowledge (semantic) is incorporated into the relations used for clustering. In the future, we will test our clustering method on another corpus of a larger size and extend our study to other variation phenomena as possible topic shifting devices.

[4] The interpretations given here are based on an oral communication with a domain information specialist.

Acknowledgements

Thanks to the reviewers for their constructive comments which I hope, helped improve this paper.

References

Bourigault D. (1994). LEXTER, un Logiciel d'Extraction Terminologique. Application à l'Acquisition des Connaissances à partir de Textes. PhD. dissertation, Ecoles des Hautes Etudes en Sciences Sociales, Paris, 352p.

Daille B. (1994). *Study and implementation of combined techniques for automatic extraction of terminology.* The Balancing Act : Combining Symbolic and Statistical Approaches to Language, Proceedings of the "Workshop of the 32nd Annual Meeting of the ACL", Las Cruces, New Mexico, USA, 9p.

David S. Plante P. (1991). *Le Progiciel TERMINO: De la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique de textes*, Proceedings of the Colloquium "Les Industries de la Langue", Montréal Nov. pp. 21-24.

Ibekwe-SanJuan F. (1997). Defining a linguistic-based methodology for tracking thematic trends in scientific publications. PhD. Dissertation, University of Stendhal, Grenoble France, 376p.

Ibekwe-SanJuan F. (1998). *A linguistic and mathematical method for mapping thematic trends from texts.* To appear in 13th European Conference on Artificial Intelligence (ECAI'98), Brighton, UK, 23-28 August 1998, pp. 170-174.

Jacquemin C. (1995). *A symbolic and surgical acquisition of terms through variation.* Workshop on "New approaches to learning for NLP", 14th International Joint Conference on Artificial Intelligence (IJCAI'95), Montréal, 8p.

Katz S.M. Justeson T.S. (1995). *Technical terminology : some linguistic properties and an algorithm for identification in text.* Journal of Natural Language Engineering, 1/1, 19p.

Smadja F. (1993). *Retrieving collocations from text : Xtract.* Computational Linguistics, 19/1, pp.143-177.

Tzoukermann E. Klavans J. Jacquemin C. (1997). *Effective use of natural language processing techniques for automatic conflation of multi-words.* SIGIR'97, 8p.

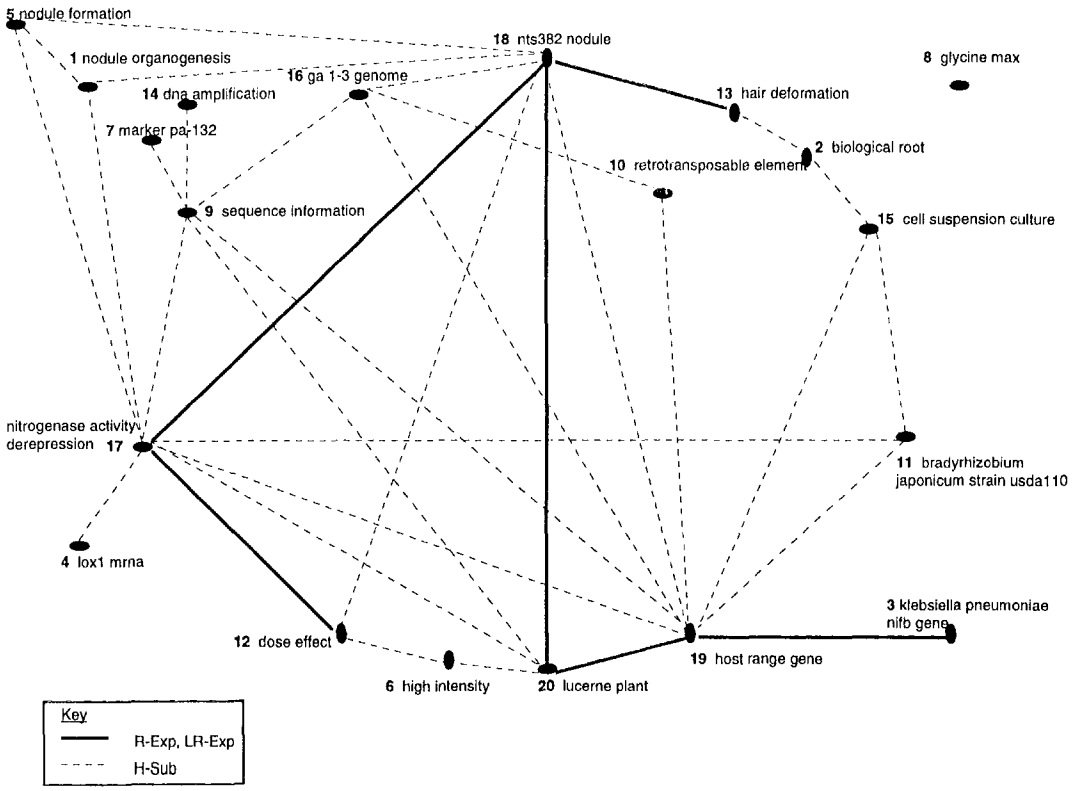

Figure 1. The external view of research topics identified in the corpus (1981-93).