Improving Automatic Indexing through Concept Combination and Term Enrichment

Christian Jacquemin*
LIMSI-CNRS
BP 133, F-91403 ORSAY Cedex, FRANCE
jacquemin@limsi.fr

Abstract

Although indexes may overlap, the output of an automatic indexer is generally presented as a flat and unstructured list of terms. Our purpose is to exploit term overlap and embedding so as to yield a substantial qualitative and quantitative improvement in automatic indexing through concept combination. The increase in the volume of indexing is 10.5% for free indexing and 52.3% for controlled indexing. The resulting structure of the indexed corpus is a partial conceptual analysis.

1 Overview

The method, proposed here for improving automatic indexing, builds partial syntactic structures by combining overlapping indexes. It is complemented by a method for term acquisition which is described in (Jacquemin, 1996). The text, thus structured, is reindexed; new indexes are produced and new candidates are discovered.

Most NLP approaches to automatic indexing concern free indexing and rely on large-scale shallow parsers with a particular concern for dependency relations (Strzalkowski, 1996). For the purpose of controlled indexing, we exploit the output of a NLP-based indexer and the structural relations between terms and variants in order to (1) enhance the coverage of the indexes, (2) incrementally build an *a posteriori* conceptual analysis of the document, and, (3) interweave controlled indexing, free indexing, and thesaurus acquisition. These 3 goals are achieved by CONPARS (CONceptual PARSer), presented in this paper and illustrated by Figure 1. CONPARS is based on the output of a part-of-speech tagger for French described in (Tzoukermann and Radev, 1997) and FASTR, a controlled indexer (Jacquemin et al., 1997). All the experiments reported in this paper are performed on data in the agricultural domain: [AGRIC] a 1.18-million word corpus, [AGROVOC] a 10,570-term controlled vocabulary, and [AGR-CAND] a 15,875-term list acquired by ACABIT (Daille, 1997) from [AGRIC].

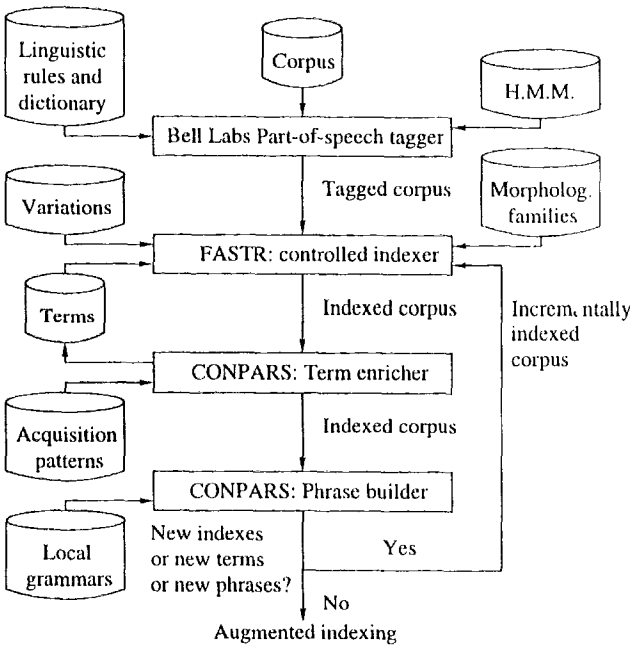


Figure 1: Overall Architecture of CONPARS

2 Basic Controlled Indexing

The preprocessing of the corpus by the tagger yields a morphologically analyzed text, with unambiguous syntactic categories. Then, the tagged corpus is automatically indexed by FASTR which retrieves occurrences of multiword terms or variants (see Table 1).

* We thank INIST-CNRS for providing us with thesauri and corpora in the agricultural domain and AFIRST for supporting this research through the SKETCHI project.

Table 1: Indexing of a Sample Sentence

La variation mensuelle de la respiration du sol et ses rapports avec l'humidité et la température du sol ont été analysées dans le sol superficiel d'une forêt tropicale. (The monthly variation of the respiration of the soil and its connections with the moisture and the temperature of the soil have been analyzed in the surface soil of a tropical forest.)

i_1	007019	*Respiration du sol*	Occurrence
	respiration du sol (respiration of the soil)		
i_2	002904	*Sol de forêt*	Embedding_2
	sol superficiel d'une forêt (surf. soil of a forest)		
i_3	012670	*Humidité du sol*	Coordination_1
	humidité et la température du sol (moisture and the temperature of the soil)		
i_4	007034	*Température du sol*	Occurrence
	température du sol (temperature of the soil)		
i_5	007035	*Analyse de sol*	VerbTransf_1
	analysées dans le sol (analyzed in the soil)		
i_6	007809	*Forêt tropicale*	Occurrence
	forêt tropicale (tropical forest)		

Each variant is obtained by generating term variations through local transformations composed of an input lexico-syntactic structure and a corresponding output transformed structure. Thus, VerbTransf_1 is a verbalization which transforms a Noun-Preposition-Noun term into a verb phrase represented by the variation pattern V_4 (Adv? (Prep? Art | Prep) A?) N_3:[1]

$$\begin{aligned} &\text{VerbTransf}_1(\ N_1\ \text{Prep}_2\ N_3\ ) \\ &\quad = V_4\ (\text{Adv}^?\ (\text{Prep}^?\ \text{Art} \mid \text{Prep})\ \text{A}^?)\ N_3 \\ &\quad \{\text{MorphFamily}(N_1) = \text{MorphFamily}(V_4)\} \end{aligned} \tag{1}$$

The constraint following the output structure states that V_4 belongs to the same morphological family as N_1, the head noun of the term. VerbTransf_1 recognizes $analysées_{[V]}$ $dans_{[Prep]}$ $le_{[Art]}$ $sol_{[N]}$ (analyzed in the soil) as a variant of $analyse_{[N]}$ $de_{[Prep]}$ $sol_{[N]}$ (soil analysis).

Six families of term variations are accounted for by our implementation for French: coordination, compounding/decompounding, term embedding, verbalization (of nouns or adjectives), nominalization (of nouns, adjectives, or verbs), and adjectivization (of nouns, adjectives, or verbs). Each index in Table 1 corresponds to a unique term; it is referenced by its identifier, its string, and a unique variation of one of the aforementioned types (or a plain occurrence).

[1]The following abbreviations are used for the categories: V = verb, N = noun, Art = article, Adv = adverb, Conj = conjunction, Prep = preposition, Punc = punctuation.

3 Conceptual Phrase Building

The indexes extracted at the preceding step are text chunks which generally build up a correct syntactic structure: verb phrases for verbalizations and, otherwise, noun phrases. When overlapping, these indexes can be combined and replaced by their head words so as to condense and structure the documents. This process is the reverse operation of the noun phrase decomposition described in (Habert et al., 1996).

The purpose of automatic indexing entails the following characteristics of indexes:

- frequently, indexes overlap or are embedded one in another (with [AGR-CAND], 35% of the indexes overlap with another one and 37% of the indexes are embedded in another one; with [AGROVOC], the rates are respectively 13% and 5%),
- generally, indexes cover only a small fraction of the parsed sentence (with [AGR-CAND], the indexes cover, on average, 15% of the surface; with [AGROVOC], the average coverage is 3%),
- generally, indexes do not correspond to maximal structures and only include part of the arguments of their head word.

Because of these characteristics, the construction of a syntactic structure from indexes is like solving a puzzle with only part of the clues, and with a certain overlap between these clues.

Text Structuring

The construction of the structure consists of the following 3 steps:

Step 1. The syntactic head of terms is determined by a simple noun phrase grammar of the language under study. For French, the following regular expression covers 98% of the term structures in the database [AGROVOC] (Mod is any adjectival modifier and the syntactic head is the noun in bold face):

$$\text{Mod}^*\ \mathbf{N}\ \text{N}^?\ (\text{Mod} \mid (\text{Prep}\ \text{Art}^?\ \text{Mod}^*\ \text{N}\ \text{N}^?\ \text{Mod}^*))^*$$

The second source of knowledge about syntactic heads is embodied in transformations. For

instance, the syntactic head of the verbalization in (1) is the verb in bold typeface.

Step 2. A partial relation between the indexes of a sentence is now defined in order to rank in priority the indexes that should be grouped first into structures (the most deeply embedded ones). This definition relies on the relative spatial positions of two indexes i and j and their syntactic heads $H(i)$ and $H(j)$:

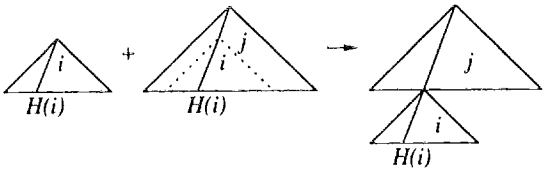
Definition 3.1 (Index priority) *Let i and j be two indexes in the same sentence. The relative priority ranking of i and j is:*

$$i \mathcal{R} j \Leftrightarrow (i = j) \vee (H(i) = H(j) \wedge i \subseteq j) \vee (H(i) \neq H(j) \wedge H(i) \in j \wedge H(j) \notin i)$$

This relation is obviously reflexive. It is neither transitive nor antisymmetric. It can, however, be shown that this relation is not cyclic for 3 elements: $i \mathcal{R} j \wedge j \mathcal{R} k \Rightarrow \neg(k \mathcal{R} i)$. (This property is not demonstrated here, due to the lack of space.)

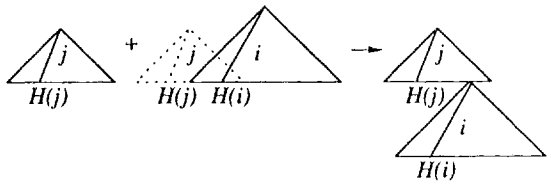
The linguistic motivations of Definition 3.1 are linked to the composite structure built at Step 3 according to the relative priorities stated by $\mathcal{R}$. We now examine, in turn, the 4 cases of term overlap:

1. Head embedding: 2 indexes i and j, with a common head word and such that i is embedded into j, build a 2-level structure:



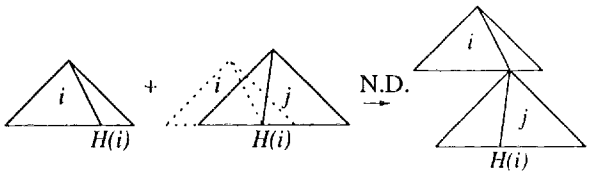
This structuring is illustrated by *nappe d'eau* (sheet of water) which combines with *nappe d'eau souterraine* (underground sheet of water) and produces the 2-level structure [[*nappe d'eau*] *souterraine*] ([underground [sheet of water]]). (Head words are underlined.) In this case, i has a higher priority than j; it corresponds to $(H(i) = H(j) \wedge i \subseteq j)$ in Definition 3.1.

2. Argument embedding: 2 indexes i and j, with different head words and such that the head word of i belongs to j and the head word of j does not belong to i, combine as follows:



This structuring is illustrated by *nappe d'eau* which combines with *eau souterraine* (underground water) and produces the structure [*nappe d'*[*eau souterraine*]] ([sheet of [underground water]]). Here, i has a higher priority than j; it corresponds to $(H(i) \neq H(j) \wedge H(i) \in j \wedge H(j) \notin i)$ in Definition 3.1.

3. Head overlap: 2 indexes i and j, with a common head word and such that i and j partially overlap, are also combined at Step 3 by making j a substructure of i. This combination is, however, nondeterministic since no priority ordering is defined between these 2 indexes. Therefore, it does not correspond to a condition in Definition 3.1.



In our experiments, this structure corresponds to only one situation: a head word with pre- and post-modifiers such as *importante activité* (intense activity) and *activivté de dégradation métabolique* (activity of metabolic degradation). With [AGR-CAND], this configuration is encountered only 27 times (.1% of the index overlaps) because premodifiers rarely build correct term occurrences in French. Premodifiers generally correspond to occasional characteristics such as size, height, rank, etc.

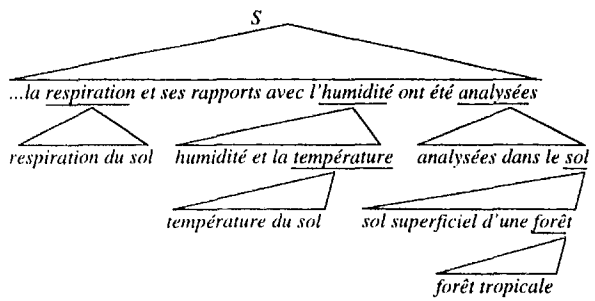
4. The remaining case of overlapping indexes with different head words and reciprocal inclusions of head words is never encountered. Its presence would undeniably denote a flaw in the calculus of head words.

Step 3. A bottom-up structure of the sentences is incrementally built by replacing indexes by trees. The indexes which are highest ranked by

the Step 2 are processed first according to the following bottom-up algorithm:

1. build a depth-1 tree whose daughter nodes are all the words in the current sentence and whose head node is S,
2. for all the indexes i in the current sentence, selected by decreasing order of priority,
   (a) mark all the the depth-1 nodes which are a lexical leaf of i or which are the head node of a tree with at least one leaf in i,
   (b) replace all the marked nodes by a unique tree whose head features are the features of $H(i)$, and whose depth-1 leaves are all the marked nodes.

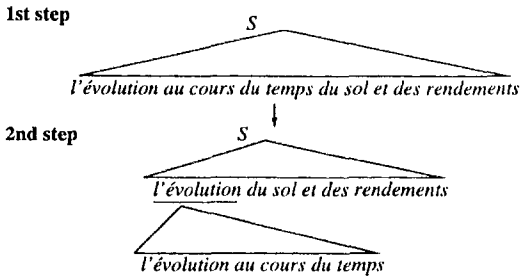
When considering the sentence given in Table 1, the ordering of the indexes after Step 2 is the following: $i_2 > i_5$, $i_6 > i_2$, and $i_4 > i_3$. (They all result from the argument embedding relation.) The algorithm yields the following structure of the sample sentence:



Text Condensation

The text structure resulting from this algorithm condenses the text and brings closer words that would otherwise remain separated by a large number of arguments or modifiers. Because of this condensation, a reindexing of the structured text yields new indexes which are not extracted at the first step.

Let us illustrate the gains from reindexing on a sample utterance: *l'évolution au cours du temps du sol et des rendements* (temporal evolution of soils and productivity). At the first step of indexing, *évolution au cours du temps* (lit. evolution over time) is recognized as a variant of *évolution dans le temps* (lit. evolution with time). At the second step of indexing, the daughter nodes of the top-most tree build the condensed text: *l'évolution du sol et des rendements* (evolution of soils and productivity):



This condensed text allows for another index extraction: *évolution du sol et des rendements*, a Coordination variant of *évolution du rendement* (evolution of productivity). This index was not visible at the first step because of the additional modifier *au cours du temps* (temporal). (Reiterated indexing is preferable to too unconstrained transformations which burden the system with spurious indexes.)

Both processes—text structuring, presented here, and term acquisition, described in (Jacquemin, 1996)—reinforce each other. On the one hand, acquisition of new terms increases the volume of indexes and thereby improves text structuring by decreasing the non-conceptual surface of the text. On the other hand, text condensation triggers the extraction of new indexes, and thereby furnishes new possibilities for the acquisition of terms.

4 Evaluation

Qualitative evaluation: The volume of indexing is characterized by the surface of the text occupied by terms or their combinations—we call it the *conceptual surface*. Figure 2 shows the distribution of the sentences in relation to their conceptual surface. For instance, in 8,449 sentences among the 62,460 sentences of [AGRIC], the indexes occupy from 20 to 30% of the surface (3rd column).

This figure indicates that the structures built from free indexing are significantly richer than those obtained from controlled indexing. The number of sentences is a decreasing exponential function of their conceptual surface (a linear function with a log scale on the y axis).

Figure 3 illustrates how the successive steps of the algorithm contribute to the final size of the incremental indexing. For each mode of

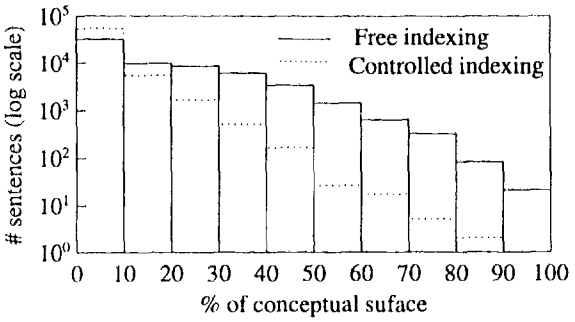


Figure 2: Conceptual Surface of Sentences

Table 2: Increase in the volume of indexing

	Acquisition	Condensation	Total
Controlled	49.3%	3.0%	52.3%
Free	5.8%	4.7%	10.5%

indexing two curves are plotted: the phrases resulting from initial indexing and from reindexing due to text condensation (circles) and the phrases due to term acquisition (asterisks). For instance, at step3, free indexing yields 309 indexes and reindexing 645. The corresponding percentages are reported in Table 2.

The indexing with the poorest initial volume (controlled indexing) is the one that benefits best from term acquisition. Thus, concept combination and term enrichment tend to compensate the deficiencies of the initial term list by extracting more knowledge from the corpus.

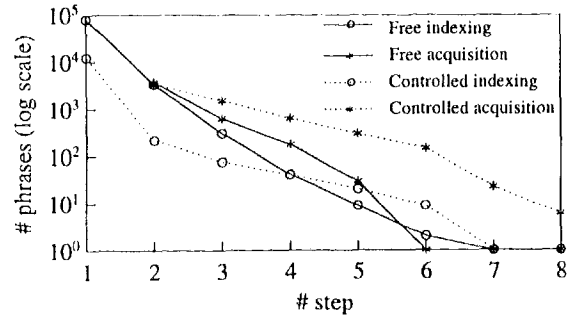


Figure 3: Step-by-step Number of Phrases

Qualitative evaluation: Table 3 indicates the number of overlapping indexes in relation to their type. It provides, for each type, the rate of success of the structuring algorithm. This evaluation results from a human scanning of 542 randomly chosen structures.

Table 3: Incremental Structure Building

	Head embedding	Argument embedding	Total
Distribution	27.0%	73.0%	100%
# correct	128	346	474
Precision	79.0%	91.1%	87.5%

5 Conclusion

This study has presented CONPARS, a tool for enhancing the output of an automatic indexer through index combination and term enrichment. Ongoing work intends to improve the interaction of indexing and acquisition through self-indexing of automatically acquired terms.

References

Béatrice Daille. 1997. Study and implementation of combined techniques for automatic extraction of terminology. In J. L. Klavans and P. Resnik, ed., *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, p. 49–66. MIT Press, Cambridge.

Benoît Habert, Elie Naulleau, and Adeline Nazarenko. 1996. Symbolic word clustering for medium size corpora. In *Proceedings of COLING'96*, p. 490–495, Copenhagen.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of ACL-EACL'97*, p. 24–31.

Christian Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In S. Wermter, E. Riloff, and G. Scheler, ed., *Connectionist, Statistical and Symbolic Approaches to Learning for NLP*, p. 425–438. Springer, Heidelberg.

Tomek Strzalkowski. 1996. Natural language information retrieval. *Information Processing & Management*, 31(3):397–417.

Evelyne Tzoukermann and Dragomir R. Radev. 1997. Use of weighted finite state transducers in part of speech tagging. In A. Kornai, ed., *Extended Finite State Models of Language*. Cambridge University Press.