Semantic Interpretation of Binominal Sequences and Information Retrieval

Cécile Fabre[1] and Pascale Sébillot[2]
[1]ERSS, Maison de la recherche, Université Toulouse-Le Mirail, 5 Allées A. Machado, 31058 Toulouse cedex, France; email: cfabre@univ-tlse2.fr
[2]IRISA, Campus de Beaulieu, 35042 Rennes cedex, France; email: sebillot@irisa.fr

Abstract

In this paper, we show, by using concrete examples from an experimentation conducted on a French system of telematic services, how a rich semantic model for binominal sequences can be used in order to increase both the recall and precision rates of an information retrieval system. More precisely, we first explain how the context of these binominal sequences, that form index descriptors of texts and requests, can be used to disambiguate nouns, provided that syntagmatic links exist or are developed within the thesaurus of the retrieval system, and that a specialized disambiguation function is used to take advantage of these relevant links; then we show that it can also permit to discover semantic paraphrase links between a question of a user of such a system and texts of an indexed database.

1 Introduction

The quality of an information retrieval (IR) system based on text indexing methods is greatly dependent on the techniques that are chosen to translate the contents of the texts into an indexing language, and to realize the matching between the indexed texts and the question of a user of the system. Two kinds of indexing methods can be distinguished, which are respectively based on the use of simple or complex indexes. In the first method, the descriptor of a text is formed from the nouns, verbs and adjectives that it contains; this technique leads to systems with a good recall rate[1], but their precision rate[2] may be low, because the indexes are too general and ambiguous [Fag87]. Complex indexing systems, currently less numerous, manipulate groups of words. Their aim is a high precision rate, but this choice may have some repercussion on the recall rate, because complex indexes are more specific, and therefore more scattered. Moreover, the extraction of complex terms is a more difficult task than collecting simple words, even if a lot of studies are developed in this domain [Jac97]. Fagan [Fag87] suggests that a middle way between these two methods can be chosen, which consists in defining *informative enough indexes*. It may correspond in a specialization of too general indexes, for example by constraining the context in which they can appear. It can also be achieved by a generalization of terms, by establishing similarity links between indexes. For example, hyperonymy or synonymy relations can be used to extend the indexes and to increase the possibilities of formulation of the same concepts. Much of current systems use a minimal reformulation of terms, based on morphological equivalences. On the other hand, conceptual indexing [Mau89] realizes matching on the base of conceptual frames.

Our work is situated in this domain of term generalization and specialization. We assume that complex nominal sequences must undergo a specific semantic treatment in order to increase the performances of IR systems. The richer the information associated with these sequences are, the greater the possibilities of manipulation and reformulation are. This idea has already been proposed by Gay and Croft [GC90], who have defined three objectives to increase the performances of text indexing, using semantic information on English compounds: determination of the conditions under which the concept expressed by a compound is present in a text, in order to increase the precision rate; recognition of equivalent reformulations (structural or semantic) of the compounds, in order to increase the recall rate, which can also be bettered by a weighting of the

[1]The proportion of relevant answers compared to the total number of relevant answers in the database.

[2]The proportion of relevant answers compared to the total number of given answers.

words of the compounds proportional to their importance. We propose to experiment their conclusions in a real application, and to complete the way of increasing the precision rate by an objective of disambiguation of polysemous words, an idea that they do not mention. Therefore, the object of this paper is to show, by using concrete examples from an experimentation conducted on a French system of telematic services, how a rich semantic model for binominal sequences can be used to increase both the recall and precision rates of an IR system. We explain the precise points that the knowledge of the semantics of complex indexes can improve; moreover, we also show that the integration of the results of a theoretical model into a real-world application has to deal with the constraints of the latter. First, we quickly present our model of interpretation of English and French binominals. Then we describe the system in which we have integrated the semantic representation of binominal sequences provided by this theoretical model. Finally, we explain the experimentation that we have conducted and show how the context of the binominal can be used to disambiguate nouns, and to discover semantic paraphrase links between the question of a user of the system and the texts of the indexed database.

2 The Domain-Independent Model of Semantic Interpretation of Binominal Sequences

In this section, we present the main ideas of the domain-independent model for the automatic interpretation of English NN and French N à/de (det) N binominal sequences that we have developed and used in our experimentation. This model is detailed in [FS95, Fab96], where its choices are justified.

Interpreting nominal compounds[3] consists in retrieving the predicative relation between their constituents. The characteristics of this model is to propose domain-independent rules of interpretation that are deduced from the morphological, syntactic and semantic features of the nominal constituents. The retrieval of the semantic link between the constituents of a compound can be a more or less problematic task, depending on the clues provided by the complex sequence itself. Two kinds of binominals are distinguished, those for which the predicate that links the two components is explicit, and those for which it is implicit. For the first class of compounds, in which at least one component is a deverbal (derived from a verb through suffixation), the model integrates results developed within the generative framework [Sel82, Lie83] and the interpretation is based on the satisfaction of the argument structure of the explicit verbal predicate. *Truck-driver* is therefore interpreted as *predicate: drive(agent: driver; theme: truck)*. The model tackles binominal sequences with no explicit predicates in the same way as the previous ones, that is, the interpretation is based on the identification of an implicit predicative information, typically associated with nouns. In order to achieve this goal, principles developed by Pustejovsky [Pus95] in his *generative lexicon* are integrated, and more precisely the *qualia structure*, which describes the relational properties of nouns. According to Pustejovsky, four kinds of predicative information are implied in the semantics of a noun: the *telic role* that refers to the purpose and function of the referent, the *agentive role* that concerns the factors involved in its origins, the *constitutive role* that captures the relation between an object and its constituent parts, and the *formal role* described as that which distinguishes the object within a larger domain. We argue that these four roles are also involved in the compounding processes, and our interpretation rules are therefore based on this conceptual description of nominals. The steps of the semantic calculus of a French or English binominal sequence are therefore: determination of the predicate(s) linked to the constituents, essentially focusing on those associated with the head noun; filtering of the predicative expressions that are really possible for the sequence - this filtering is based on semantic type constraints on the arguments of the predicates and on the semantic role of the preposition and the determiner in French. To obtain this semantic information, the WordNet lexical database is used.

This theoretical model presents a rich representation for both French and English nouns, and therefore lots of possibilities for interpreting complex sequences in these two languages (cf. [Fab96] for a valuation of its performances). Moreover, it is domain-independent and may therefore be used in applications where texts range over a great number of topics. We have integrated its principles within an IR system, and have conducted an experimentation in order to valuate the repercussion of this rich representation on such a system. This study

[3] In this paper, we use equally the words *binominal sequence*, *binominal*, *compound* or *complex sequence* for English NN and French N à/de (det) N sequences.

was supported by a French telecommunication center (CNET), and we have had to deal with the precise constraints of its system; so we begin the presentation of the experimentation by a quick description of the system.

3 The CNET Information Retrieval System

The CNET IR system has been developed for a natural language questioning of a database of French telematic services, and is proposed on the Minitel system. The texts that describe the services are short, generally composed of three to five lines [GLO93] (eg. 3617omega: "*Information boursière. liste de marchés, cours, signaux*" (Stock exchange information. market list, quotations, signals); 3617actor: "*Saisie de CV pour acteurs. recherche d'acteurs*" (CV input for actors. research of actors)). In Fagan's [Fag87] and Mauldin's [Mau89] terminology, the CNET system can be described by the five following characteristics: preindexing of the texts with a reverse file method, vectorial model for text matching [Sal75], atomic indexes, minimal linguistic analysis of the texts and use of a thesaurus. It is composed of three modules: the linguistic analyser, the indexing module and the matching module. The *linguistic analyser* generates a structured representation of a text. Each sentence undergoes a lexical analysis and a syntactic parsing, which determines its syntagms but does not link them to each other. Then a semantic representation is produced in which each word is replaced by the list of its meanings, found in a lexicon (*word_x* indicates the meaning number *x* of the word *word*). The *indexing module* generates the list of the indexes that represent the contents of a text. The list of all the notions produced during the semantic analysis is completed with semantically related indexes through a process described later. The indexes are weighted according to their frequencies in the text and in the database. The *matching module* valuates the relevance of a text for a given question, by computing the cosine of the angle between the question and the text vectors. The system integrates a semantic network (thesaurus) constituted from a French dictionary. The nodes corresponds to the meanings of the words. There are two kinds of arcs: the "is-a" arcs (noted superc) and the "relating-to" arcs (noted relatif-à) which correspond to different kinds of links like paraphrase (paraph), synonymy (syn), domain (dom) or has-in-its-definition (a-dans-def). The network is consulted twice. The linguistic analyser possesses a disambiguation function which searches a path in the semantic network between the different meanings of the words that have been identified during the semantic analysis; only the meanings linked in the network are kept. During the indexing phase, the thesaurus is used to add notions that are semantically close to the list of the indexes corresponding to the words found in the text (or question). The arcs coming out of the senses produced by the semantic analysis are followed up and the senses that are found are added to the list; this process is controlled in order to avoid the addition of too distant notions: following up a link divides by two the weight of the initial meaning and the *propagation* is bound by a *limit*. If a word has n meanings in the initial list of indexes, the weight of each initial meaning is 1/n. The database of texts is preindexed; a valuation of the system on 10 questions [GLO93] has led to a 62.8% precision rate and a 33.4% recall rate.

4 The Experimentation

4.1 First Tests

Our experimentation has been conducted on a database of French texts that corresponds to 166 Minitel "3617" services (cf. 3617omega or 3617actor examples above). The first step has consisted in testing the system with N à/de (det) N questions, in order to precisely measure the repercussion of a semantic treatment of these sequences on the results of the indexing. We have produced a corpus of 50 question/answers pairs. If we consider the 5 highest weighted services proposed for each question, the precision rate is about 40% and the recall rate about 30% (respectively 80% and 30% if all the answers are considered). Two problems related to the semantic of binominals have appeared: an insufficient disambiguation and an insufficient exploitation of the semantic information beared by the sequence.

If a question contains a polysemous word, the disambiguation does not work and all the meanings of the word are used to find texts. For example, the question "*cours de dessin*" (drawing course) has led, with the highest scores, to services about change rates (*cours de monnaie* in French). We have explored the semantic network to understand the problem and seen that it was quite incomplete. We have added some links between elements like ("cours_1221", "monnaie_1" : a-dans-def), accord-

ing to definitions in the dictionary that had not been integrated into the thesaurus. Even with these new links, the results remain identical. This is due to the weightings: *cours* has 10 meanings, so each meaning gets an initial 1/10 weight; therefore every one-step propagation in the network leads to a 1/20 weighted meaning, which is below the propagation limit.

The second problem has a big influence on the precision rate, but also on the recall rate. Indexing through simple indexes precludes the correct consideration of the complex concept corresponding to the sequence. For N_1 prep N_2 questions, the system firstly gives texts with N_1 *and* N_2 indexes, then with N_1 and prop(N_2) or conversely, where prop(X) is an index propagation from X, then N_1 *or* N_2 indexes, and finally prop(N_1) or prop(N_2) indexes. In a lot of answers, because of the propagation limit, only texts containing either N_1 *and* N_2 indexes or just one N_i index are given; the notion of conceptual similarity is not taken into account.

4.2 Objectives

The link between the theoretical model and this application consists in using the semantics of binominal sequences to disambiguate the constituents and to determine similarity links between concepts in the question and in the texts.

Concerning the production of indexes, the idea is to take advantage of the context of a binominal to improve the disambiguation of its constituents. In the theoretical model, a polysemous constituent may be disambiguated in two cases: if it is identified as an argument of a predicative expression associated with the second element, only the meanings compatible with the selectional restrictions on this argument are kept; if it is associated with several predicative expressions, only those that furnish an argumental role to the second constituent are kept. The principles of constituent disambiguation of this model are therefore quite different from those governing the disambiguation in the CNET IR system, which are based on *paradigmatic* links, that is, where the question is to identify synonym or cohyponym concepts. In the model, the links are *syntagmatic* ones; the question is to determinate which meaning of the head is compatible with which meaning of the modifier, and these associations involve the predicative roles that are defined in the model. This kind of links is virtually missing in the CNET thesaurus. Therefore our proposition to tackle the disambiguation problem is to develop semantic links based on the nominal modification in the network, and to define a new disambiguation function that explores these relevant links.

Concerning the matching module, the tests have shown that silence is essentially caused by an insufficient account of the equivalence between distinct formulations of the same idea. The reformulation may be a morphological or syntactic paraphrase, but also a semantic one. This case corresponds to the replacement of one constituent by a semantically close concept (synonym, hyperonym, etc.), but also to the paraphrase of the predicate of the relation expressed by the sequence (eg. *magasins de disques* (disk store) and *vente de disques* (disk sale)). The identification of the relation between the constituents of the binominal realized in the theoretical model enables to exhibit a predicative information that can be exploited in the semantic paraphrase links. It extends Fagan's hypothesis [Fag87] which brings in the distinction between paradigmatic and syntagmatic dimensions that we have defined within the disambiguation framework. According to Fagan, the identification of *thesaurus relations* (synonymy, etc.) enables to increase recall rate, whereas *phrase relations*, which are modification or specification relations within nominal or verbal syntagms, influence the precision rate. The semantic calculus for binominals enables us to access to the syntagmatic dimension, by identifying a source of variation from predicative information. During our preliminary tests, we have found this kind of (unresolved) paraphrases between a nominal and a verbal sequences (*création d'entreprise; créer son entreprise*) (company creation; to create a company), two nominal sequences whose heads are associated with the same predicative information (*cours de maths; professeur de maths*) (maths lesson; maths professor), two nominal sequences whose modifiers are associated with the same predicative information (*matériel de navigation; matériel pour bateaux*) (navigation equipment; boat equipment). However, considering the constraints of the CNET system (no links between the syntagms during the syntactic parsing, production of atomic indexes (no information of cooccurrence, of order, or proximity), a preindexed database of texts (we were not allowed to re-index it, and have only work on the production of new indexes for the questions), and an incomplete thesaurus (cf. above)), it has been impossible for us to test all our ideas about semantic reformulations of concepts, for example the correspondence between Noun Verb and Noun prep Noun formulations. We have only been able to work on the reordering of

the proposed answers; however we have lowered the propagation limit in order to obtain more answers, with a new control of their quality. Concerning this point, we have tested the following hypothesis: "if there exists a *nominal modification link* between two concepts A and B, always prefer texts containing the double index prop(A) prop(B), where prop(X) corresponds either to the concept X or to a concept obtained from the propagation of X through the network". Here are the results of the experimentation.

4.3 Results

A specialized disambiguation function for the indexing module: In our corpus of tests, we have selected the cases where the quality of the answers was affected by a disambiguation failure. For these 7 cases, we have added links in the semantic network, named "arg", that illustrate the different kinds of predicative expressions found in the theoretical model: *verbal subcategorization*: "traitement_2", "maladie_1" : arg (a treatment in its medical meaning treats diseases); *agentive relation*: "émission_311", "télévision_2" : arg (a broadcast is broadcasted by a television). The manual addition, based on the definitions in the dictionary, is a hard task. The definitions are sometimes given with different words, therefore it is impossible to do a systematic work; moreover the links in the network bind disambiguated concepts whereas definitions manipulate words; the good meanings that must be linked have to be found.

The previous disambiguation function, which firstly initializes the concepts that must be propagated, then propagates through the *is-a* links and finally through the *relating-to* links, has been replaced by a specialized function that firstly initializes the concepts that must be propagated, then propagates *the head of the sequence* through its *arg* links and its paradigmatic links (*syn, is-a*), and finally propagates *the modifier* through its paradigmatic links. This new function enables us to treat correctly the 7 cases that were previously not disambiguated. For example, if we do not add "arg" links and keep the first disambiguation function, "*cours de dessin*" (drawing course) leads to 10*5 possible interpretations (10 meanings for *cours* and 5 for *dessin*); if we add the links but keep the previous function, it leads to 10*1 interpretations; if we both add the links and use the new disambiguation function, it only leads to 1*1 interpretation, which is the correct one.

Semantic reformulation - Reordering of answers: Since we were only allowed to act after the production of the answers, our role here has consisted in trying to reorder them correctly, by distinguishing those containing the complex notion corresponding to the binominal sequence in the question, according to the principles described in the preceding section. First, we have modified the production of the index vector of the question, in order to recognize which concepts are present in the question and which ones are obtained through propagation. Secondly, we have added some paradigmatic links in the incomplete semantic network, according to the definitions of the concepts in the dictionary, and have lowered the propagation limit, in order to have an opportunity to increase the recall rate. After these changes, we have tested our hypothesis on the 50 elements of our corpus.

The results, as shown in Table 1, do not seem to improve the retrieval performance in a significant way. Yet, if we make a closer look at the answers, we see that wrong reorderings are mainly due to the fact that the system relies on a thesaurus in which many semantic connections are both fragmentary and unappropriate.

For 32 questions (3/5), no reordering was observed. 21 cases correspond to the fact that no text in the database contains the cooccurrence of the concepts; 11 cases correspond to the fact that the texts with the cooccurrence were already classified in the first places. However, in these last cases, we can distinguish two subcases, which correspond respectively to the fact that the detected cooccurrence concerns services whose weights are really higher than those of the other services (9), or concerns services whose weights do not permit to distinguish them from the others (2); for these last cases, the detection of the cooccurrence permits to make a clear distinction with the following services.

The 18 last cases (2/5) show a reordering of the answers. In 5 cases, the reordering is correct and really improves the quality of the answers that are proposed (higher weights and upper places for the most relevant services). For example, for the question "*industrie du textile*" (textile industry), the 3617infotextil service: "*Informations sur le secteur cuir, textile, habillement*" (Informations about the leather, textile and clothes areas) is replaced before 3617fdi: "*Industrie Française de Défense. entreprises, matériels d'armement*" (French defense industry. companies, arming equipment) and before 3617chimitel: "*Banque de données de fournisseurs dans l'industrie chimique*" (Database of suppliers in

No reordering	Good reordering	Wrong reordering	Partly good / partly wrong reordering
32	5	7	6

Table 1: Proportion of reorderings on the corpus

the chemistry industry). Two other examples are shown in Table 2 and Table 3.

Text	Reordering
"Annonces d'emploi pour cadres et commerciaux, calcul de salaires." (Jobs for executives, wages calculations)	1 → 1
"Informations pour les cadres des entreprises. rémunérations." (Information for executives. wages)	5 → 2
"Consultation d'offres d'emploi. cadres et non cadres." (Advertisements for jobs. executives and non-executives)	2 → 3
"Mobilité du travail en France et salaires moyens par région." (Wages in France)	equal 3 → equal 4
"Evaluation du salaire selon la fonction" (Estimation of wages)	equal 3 → equal 4

Table 2: Reordering of the answers to the request *salaires des cadres* (*wages of executives*)

Seven cases correspond to wrong reorderings and 6 to both right and wrong ones inside the list of answers. The wrong reorderings have two causes: first, problems due to the network appear; some services are prefered because of some incongruous semantic links (eg. "carburant_2" (fuel) is linked to the adverb "officiellement" (officially) by a "a-dans-def" (has-in-its-def) link!); secondly, problems due to a wrong disambiguation are stressed. For example, the question "*métiers du sport*" (sport professions) leads to the service "*Laboratoire Cami, sélection de formules pour machines à teinter*" (Cami laboratory, selection of formula for tinting machines) because of the double link between *métier* in the meaning weaving machine and *machine*, and between *formule* and *sport* (the meaning formula_1 being obtained through propagation from formula).

Text	Reordering
"Les annonces du Parisien Libéré" "emploi, auto, moto, immobilier" (Advertisements for jobs, cars, etc.)	equal 2 → equal 1
"Annonces auto, prix du neuf, cote de l'occasion." (Advertisements for cars)	equal 2 → equal 1
"Centre Communautaire des Particuliers" "import de voitures" (Car imports)	1 → 3
"Annonces d'emploi à l'étranger" (Advertisements for jobs)	4 → 4

Table 3: Reordering of the answers to the request *annonces de voitures* (*advertisements for cars*)

5 Conclusions and Future Works

The experimentation allows us to conclude that a rich representation of the semantic of binominal sequences improves the recall and precision rates of an IR system. The context of the binominal can be used to disambiguate nouns provided that syntagmatic links are developed within the thesaurus, and that a specialized disambiguation function is used to take advantage of these relevant links. It can also permit to discover semantic paraphrase links between a question and texts, and we have shown that the matching must favor the texts containing the complex concept expressed by a sequence compared with those containing only a concept associated with one of its constituents.

The experimentation shows that retrieval performance (and especially precision rate) can be improved providing that the system relies on a rich semantic thesaurus; the CNET thesaurus, which displays many unappropriate semantic connections, has limited the reordering of the answers, so that false semantic relations have emerged. Therefore, our first future work concerns the necessity to possess a dense

and correct thesaurus, and we are currently working on the automatic acquisition of nominal modification links from corpora. Moreover, the technical impossibility, in the described experimentation, to get syntactic links between the elements of a text has limited our first objectives. Particularly, the second part of the experimentation has not treated the possibilities of reformulations of the whole sequences, but only that of their constituents. Our second aim concerns this point, and more generally, the study of the semantic variations of terms in the same way as works developed by Jacquemin [Jac97] on their morpho-syntactic variations.

References

[Fab96] C. Fabre. *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations.* PhD thesis, Rennes 1 University, France, 1996.

[Fag87] J.L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: a Comparison of Syntactic and Non-Syntactic Methods.* PhD thesis, Cornell University, Ithaca, New-York, USA, 1987.

[FS95] C. Fabre and P. Sébillot. Calculability of the Semantics of English Nominal Compounds: Combining General Linguistic Rules and Corpus-Based Semantic Information. Research Report n^o2742, INRIA, 1995.

[GC90] L.S. Gay and W.B. Croft. Interpreting Nominal Compounds for Information Retrieval. *Information Processing and Management*, 26(1):21–38, 1990.

[GLO93] M. Gilloux, E. Lassalle, and J.-M. Ombrouck. Interrogation en langage naturel du Minitel Guide des Services. *Écho des Recherches*, 1993.

[Jac97] C. Jacquemin. Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Habilitation à diriger des recherches, Nantes University, France, 1997.

[Lie83] R. Lieber. Argument Linking and Compounds in English. *Linguistic Inquiry*, 2(14):251–285, 1983.

[Mau89] M.L. Mauldin. *Information Retrieval by Text Skimming.* PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 1989.

[Pus95] J. Pustejovsky. *The Generative Lexicon.* MIT Press, 1995.

[Sal75] G. Salton. *A Theory of Indexing.* Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[Sel82] E. Selkirk. The Syntax of Words. *MIT Press*, 1982.