A Corpus-Based Approach to Refining a Domain-Independent Compound Interpretation System

Pascale Sébillot
IRISA
Campus de Beaulieu
F-35042 RENNES Cedex, FRANCE
April 5, 2000

Abstract

In this paper, we present a domain-independent model for the automatic interpretation of nominal compounds in English and in French. This model is meant to account for productive rules of interpretation, that are inferred from the morpho-syntactic and semantic characteristics of the nominal constituents. We then propose the corpus-based approach that has been used to adapt this general model to the interpretation of compounds of a specific domain.

1 Introduction

Nominal compounds represent a very large part of the terminology that can be extracted from technical texts. Their automatic interpretation is therefore fundamental in a lot of applications of natural language processing (automatic translation, text indexing, etc.). Works on automatic interpretation of compounds in a specific domain have led to systems [14, 22, 25] that are quite efficient. However, since these systems are developed within a restricted domain, they are based on very specialized lexical information, and it is very difficult to generalize them outside the domain they are built for. Some general systems [10, 4], built to account for any kinds of interpretation patterns, also exist, but they contain rules that are not inferred from properties of the constituents of the compounds, and therefore weighted by frequency and probability scores hardly defensible in the absence of any reference to a domain.

In order to avoid both these problems, we\textsuperscript{1} have developed a domain-independent system of compound interpretation [7], based on non-specialized lexical information, that accounts for productive rules of interpretation that are

\textsuperscript{1}This work is developed together with C. Fabre and R. Pichon at IRISA.
inferred from the morphosyntactic and semantic characteristics of the nominal constituents. This system focuses on the calculability of the semantics of both French and English compounds, that is, it automatically retrieves the semantic relation between the constituents in \( N \) \( N \) (Noun Noun) sequences in English (e.g., window manager) and \( N \) à/\( de \) \( N \) sequences in French (e.g., gestionnaire de fenêtres). It has been tested [7, 8] on sequences randomly picked up from large corpus [24] for English, and [16] and [3] for French, disconnected from their original contexts. The results are quite correct and the system generates all the meanings that are possible, considering the constituents.

However if we want to apply this general system to a given domain, the multigeneration of solutions has to be limited to only produce the meanings that are correct in the considered context. Moreover some solutions that are produced by the general interpretation system have to be refined in a particular context. In order to achieve these goals, we have developed a corpus-based method that grasps specific contextual information from the texts of a domain, and adapts the domain-independent model to build a more accurate one. The interest of our work over previous published ones is that we have a general system of interpretation, that can therefore be used in a lot of domains, and a method to adapt this system to a given context. Among others, this method allows us to define preferential rules of interpretation among possible rules for a compound.

In this paper, we first briefly present the principles of the general model of interpretation of compounds. We then explain what kind of information can be obtained from a corpus in order to adapt the model to a given domain.

2 The Domain-Independent System

Interpreting nominal compounds consists in retrieving the semantic relation between the constituents, on the grounds of their morphological, syntactic and semantic characteristics. We only present here interpretation examples of English \( N \) \( N \) compounds, but the same model can also be used for French \( N \) à/\( de \) \( N \) sequences.

The retrieval of the semantic link (or predicate) between the constituents of a compound can be a more or less problematic task, depending on the clues provided by the compound itself. We distinguish two kinds of compounds, those for which the predicate that links the two components is explicit, and those for which the predicate is implicit.

2.1 Compounds in which the Predicate is Explicit

These compounds are characterized by the fact that the predicate that represents the semantic relation between the constituents is explicitly linked to one of the components, on morphological, syntactical and semantic grounds. One of the constituents is therefore a deverbal noun, resulting from the suffixation of a verbal base, and the predicate corresponds to this verbal root.
To tackle with compounds of that type (e.g. parsing program), our model integrates results developed within the generative framework [23, 12]. The interpretation is based on the notion of satisfaction of the argument structure of the explicit predicate. We distinguish two types of deverbals:

- **action deverbals**, that refer to the accomplishment or the result of the process referred to by the verb (e.g. parsing). These deverbals inherit the entire argument structure of the verb (e.g. parse(agent, theme) \(\rightarrow\) parsing(agent, theme)).

- **subject deverbals**, that refer to one actor of the process referred to by the verb (agent, instrument, e.g. generator, or patient, e.g. employee). These deverbals inherit the argument structure of the verb deprived of the argument controlled by the suffix (e.g. parse(agent, theme) \(\rightarrow\) parser(theme)).

When the deverbal noun fills the head of the compound, the modifier (i.e. generally the leftmost constituent in English) may satisfy one of the arguments mentioned in the argument structure of the deverbal and fill a thematic role: sentence parsing \(\rightarrow\) predicate: parse(theme: sentence)\(^2\), or it may fill a semantic role, referring to a circumstance of the action (location, time, means, etc.): hand parsing \(\rightarrow\) predicate: parse(means: hand).

When the deverbal noun is on the left of the compound (i.e. when it occupies the modifier position), it cannot satisfy its object argument within the compound; in this case, the head may only fill a semantic or agentive role: parsing program \(\rightarrow\) predicate: parse(agent: program). For a detailed account of these principles, see [21].

### 2.2 Compounds in which the Predicate is Implicit

We try to tackle with these compounds in the same way as the previous ones, that is, we try to find one (or more) predicate(s) associated with the simple nouns that form the compounds. In order to achieve this goal, we integrate in our model principles that J. Pustejovsky [17] has developed in his generative lexicon, and more precisely the *qualia structure* which describes predicative attributes associated with a noun in its lexical definition. For 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\(^3\), 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 (see [7, 9] for a detailed account of it). The description allows us to differentiate the interpretations of pill in seasickness pill \(\rightarrow\) predicate:

\(^2\)For a detailed account of the semantic representation that is chosen, see [6].

\(^3\)This attribute corresponds to the notion of *role nominal* in [10].
healing (instrument: pill, theme: seasickness) (telic relation), and in antihistamine pill → predicate: made of (object: pill, source: antihistamine) (constitutive relation).

The determination of the type of role involved in a given compound is linked to the semantic class of the second constituent. Therefore we use the WordNet lexical base, which proposes a hierarchy of semantic classes to which the lexical entries of our interpretation system are linked. We only use the upper part of the hierarchy in order to remain domain-independent. Predicates that are specific for the interpretation of a noun, particularly when it is associated with a second constituent in a compound, are linked to this single noun. For example, a predicate like cut is linked to a noun like knife, but cannot be generalized to the semantic class instrumentality to which knife belongs. More generic predicates that permit the interpretation of an entire semantic class of nouns (e.g. contain or hold for container) are associated with this class in the hierarchy.

We now present briefly the results of this domain-independent system, and we explain what must be improved, and how, if we want to use it in a specific domain.

3 Application to a Given Domain: a Corpus-Based Approach

We first point out some of the results of the tests of the previously described system. We then explain how the use of a corpus has helped us to apply this domain-independent system to a given domain.

3.1 Results

Our modeling of the semantics of nominal compounds has been implemented and tested on a list of sequences randomly picked up from large corpora: R. Sproat’s [24] one for English compounds, and A. Poncet-Montange’s [16] and B. Daille’s [3] ones for French compounds. We do not describe here the results in general (for a detailed account of it, see [7] and [8]), but we point out elements that are characteristic of the general model and that must be improved when the system is applied to a precise domain.

Multigeneration of Solutions: In our general system, different rules can be applied for a same compound, for example because the head noun may both be a deverbal and belong to a given semantic class. Moreover, a given noun which belongs to a specific semantic class inherits the properties of the ascendant classes in the hierarchy of WordNet, and if a predicate is associated with each level of the hierarchy, different solutions are produced.

There are in fact two kinds of multigeneration.

---

4WordNet is a trademark of Princeton University. Cf. [15].
• The first one can be illustrated by the three interpretations that the system gives for the compound wood-checker:
  → predicate: check(agent: checker, theme: wood),
  → predicate: check(instrument: checker, theme: wood),

The suffix -er associated with check leads to two interpretations of the deverbal checker, one instance with the feature person, and the other one with the feature instrumentality. The third meaning comes from the fact that the feature instrumentality is attributed by a predicate made of in the hierarchy, predicate that can be instantiated by a modifier of the class substance like wood. The three solutions are correct outside any context, but in a precise domain only one is the correct interpretation of the compound.

• The second kind of multigeneration can be illustrated by the interpretations of liquor bottle:
  → predicate: hold(instrument: bottle, theme: liquor),
  → predicate: made of(object: bottle, source: liquor).

The first interpretation is generated because bottle belongs to the class container to which the predicate hold is linked. But in the hierarchy, the class container is a descendant of the class artifact to which the predicate made of is linked. So the system gives two interpretations, one which is correct and the second which is surely mistaken (but which is a correct pattern of interpretation for paper bag for example), but which can only be filtered by extralinguistic knowledge. The only way to avoid this second interpretation is to have strict constraints on the types of elements that can be a substance for a bottle, information that cannot be put in a general system.

Not Precise Enough Solutions: A second problem of the system, if we apply it directly to a given domain, is that the predicates that are associated with classes (or even with nouns) may be too general in that domain.

No Solutions: Some compounds have obtained no interpretation during the tests. For example, when the link between the constituents is a sort of or a resemblance one (e.g. C language, cucumber fish), the system fails in giving any meaning. We do think that these kinds of links must be learnt from a corpus.

3.2 Corpus Help to Apply the System to a Given Domain

A corpus is a collection of texts that are representative of a language (and consequently, of the language of a precise domain). These texts may therefore be considered as the production set of the rules that govern the behaviour of
that language, and a lot of language aspects that are hardly formalizable can be inferred from their manifestations in the corpus. Corpora are a very interesting source of information in numerous natural language processing applications, and are used for two main tasks: model validation, when the corpus is used to test the pertinence of an already elaborated model of the behaviour of the language (see for example [2, 13, 27]), and hypothesis validation (or information acquisition), when a rule of the language is not well formalized and is tested with the help of the corpus (see for example [11, 26, 1]). We present here the way we have analysed and used a corpus of a given domain in order to adapt our domain-independent model of compound interpretation to this particular domain. The method that we describe is a systematic one, that can be used to adapt the general model to different types of corpus and domains.

In our tests, we use two different corpora: R. Sproat’s one [24] and the British National Corpus (BNC\(^5\)). The first corpus is constituted of compounds coming from newspapers, especially from socio-economic and political articles. The second one contains tagged texts of different domains. We choose various parts of these corpora in order to test the adaptation of the model to different domains.

For presentation purpose, we split the elements obtained through the corpus analysis in four parts, even if the frontiers between these parts are not strict ones.

**Acquisition of General Information on the Compounds of a Specific Domain:** The first use of the corpus is a characterization of the compounds of the precise domain. This characterization firstly concerns the syntactic form of the compounds. For example, we learn if the \(NN\) compounds are preferentially formed with simple nouns, or if they contain deverbals. In the case of deverbal nouns, we identify the most often used suffixes. The same tests have been done on French corpus in order to know if the compounds are preferentially \(NN\) or \(Nd\) \(eN\) ones, and even if a determiner is used or not before the second noun (\(Nd\) \(de\) \(aN\) for example).

We also determine what are the most productive patterns of compounding in the domain; we may expect that the new compounds of this domain will follow these detected interpretation patterns. More precisely, we also focus on the determination of the most frequent associations of semantic classes in the texts. Like Velardi et al [26, 1] who calculate preferential concept association in a given syntactic relation (e.g. \(N_1\)-prep-\(N_2\)) in order to use these results later to desambiguate a structure, our calculus is built on frequency and conditional probability determinations.

More precisely, our method is grounded on the notion of *sense tree*. In our lexicon, each noun is linked to one or more hierarchies of semantic classes. For example, *window* may have the four following meanings, and therefore be attached to the four following WordNet class hierarchies:

- a window in a wall

\(^5\)The BNC is distributed by Oxford University Computing Services.
The sense tree of a noun is the merging of the different meaning hierarchies of this noun, that is, a tree in which each node represents a semantic class and the number of times where this class is encountered in the different meanings of the noun. For example, the sense tree of window contains the node (opening, 2), the node (artifact, 4), etc.

Given a corpus of a domain, we first build reference sense trees for all the modifier nouns and all the head nouns of the compounds in this corpus. Then, we associate with each of the most frequent modifiers the sense tree corresponding to the merging of all the sense trees of the nouns that are heads in compounds in which the left constituent is the considered modifier. Lastly, using the reference sense trees of the domain, we calculate the strength of the link between the different semantic classes of the modifiers and the semantic classes of the heads.

These first results are particularly interesting to interpret never-encountered compounds of the domain, for which no clues of interpretation are available.

Identification of Relations that can only be Detected within a Corpus:
Some links between constituents of compounds do not depend on the properties of the constituents. For example, the resemblance relation between cucumber and fish in cucumber-fish is part of extralinguistic knowledge. P. Downing [5] and M.E. Ryder [20] have defined statistical patterns of interpretation for these kinds of compounds. For example, Downing’s article is based on a survey made of a number of English-speaking subjects in order to determine the types of semantic relationships which are possible between the terms of a NN compound. Twelve types of semantic relationships were discovered: whole-part (e.g. duck foot), part-whole (e.g. pendulum-clock), composition (e.g. stone furniture), etc. The semantic relationship between the two terms depends largely on the semantic class of the head noun. Five main classes exist in Downing’s opinion: human, animal, plant, natural object and synthetic object. For each semantic class of the head-noun, she gives the relationships which may hold between the two constituents of a compound, by decreasing order of probability. For example, if the head-noun is a natural object, the relation is preferably composition, then origin, and at last place. We argue that so specific information cannot be coded in a general model of interpretation, both because the relation cannot be
retrieved on the basis of semantic knowledge on the constituents and because their statistical results highly depend on the corpus that have been used to compute them. Following J. Pustejovsky et al [18], we use mutual information measures in order to extract relations between words like the taxonomic one in C language or the resemblance one in cucumber-fish.

**Determination of (precise) Interpretation:** In the presentation of the results of the tests of the general model, we have pointed that it is not obvious to get a predicate for each semantic facet (telic, agentive, etc.) of a noun, or to get a precise enough predicate, a priori. A corpus often gives a solution to this problem. For example, when no link, or a too general one, can be obtained between two elements of a compound, we have determined a predicate by studying, in a window of n words in sentences of the texts of the corpus, which verbs link the two concerned nouns, or even two nouns of the same semantic classes. Another solution to the problem is to study the relational scheme of the head noun within the corpus.

**Precise Adaptation to the Domain:** The study of the predicates that link the members of compounds in a sentence can also be used to refine the predicate associated with nouns or semantic classes. The same kind of analysis of the corpus can also discriminate between the multigenerated solutions of the general model. For example, we can discriminate between the three interpretations of wood-checker by looking at the way checker is used in the corpus (as a human-being or an instrument, and to check wood or made of wood). We also use selectional association measures defined by P. Resnik [19] to discriminate between the multi-solutions.

**4 Conclusions and Future Works**

In this paper, we have firstly presented a domain-independent system of automatic compound interpretation. We have described how a corpus-based approach permits to adapt this general system to any given domain. All the problems mentioned if we directly use the general system to interpret compounds of a specific domain have been reduced with the help of this corpus-based approach.

Our approach is purely statistical; we use frequency calculus, mutual information and selectional association measures. Our future works will concern the comparison of the results obtained by the adaptation of the general model to a specific domain using this statistical method to those that can be obtained through the use of an automatic learning method; more precisely, by the use of a CART (classification and regression trees) method.
References


