Interpretation of Nominal Compounds: Combining Domain-Independent and Domain-Specific Information

Cécile Fabre
IRISA
Campus de Beaulieu
35200 Rennes
France
cfabre@irisa.fr

Abstract

A domain independent model is proposed for the automated interpretation of nominal compounds in English. This model is meant to account for productive rules of interpretation which are inferred from the morpho-syntactic and semantic characteristics of the nominal constituents. In particular, we make extensive use of Pustejovsky's principles concerning the predicative information associated with nominals. We argue that it is necessary to draw a line between generalizable semantic principles and domain-specific semantic information. We explain this distinction and we show how this model may be applied to the interpretation of compounds in real texts, provided that complementary semantic information are retrieved.

1 Motivation

Interpreting nominal compounds consists in retrieving the predicative relation between the constituents. In many cases, no surface information is available to deduce the relation, and in particular no morphological evidence of a link between the constituents and the underlying predicate. This problem has been tackled in several types of NLP systems, mainly:

- domain-dependent systems. Such systems are very efficient but are limited to the domain they are built for: interpretation rules are inferred from the observation of specific semantic patterns (Marsh, 1984) or from a fine-grained conceptual representation (Ter Stal, 1996).

- domain-independent systems (Finin, 1980; Mac Donald, 1982), built to account for any kind of interpretation patterns, including rules that are not inferred from the properties of the constituents (what Finin calls *productive rules*, in opposition to *structural rules*). Frequency and probability scores are added to the rules. Such numeric weighting of general semantic rules is hardly defensible in the absence of any reference to a domain.

Consequently, the questions that we propose to answer are: how far can we go in designing a model of interpretation rules which account for productive patterns of interpretation, independently of any domain? Conversely, what domain-specific information must be available to enrich this general model? The aim of our research is to define as precisely as possible the border line between what can be regularly described with general linguistic mechanisms, and what has to do with subregular or irregular phenomena which depend on corpus characteristics. This is a crucial issue when dealing with compound semantics because regular semantic patterns (involving relational properties of nominals) and extralinguistic data are mingled.

We have designed a model[1] that accounts for *structural rules* (in Finin's terminology) of interpretation of N N compounds[2], i.e. domain-independent rules that are deduced from the morpho-syntactic and semantic characteristics of the nominal constituents. The interest of this general model is to base the interpretation of compounds exclusively on general principles regarding the association between nouns and predicative information. Besides, this non-specialized model of interpretation allows us to draw a comparison with nominal sequences across languages, and es-

[1] This project is supported by the CNET (contract CNET-INRIA n°951B030). Our model of interpretation of nominal compounds will be used to enrich information retrieval in a system that is open-domain.

[2] In this work, we only focus on non-recursive terms. The same interpretation mechanisms can be extended to compounds with three constituents or more, but furthermore these compounds raise the problem of ambiguous bracketing (Resnik, 1993).

pecially with French sequences of the form "N de N" and "N à N", in which the prepositional link is semantically weak (Fabre and Sébillot, 1994).

We first describe this model, showing how compound interpretation must rely on an accurate description of the predicative properties of nominal constituents. We then suggest how this general model may be applied to the interpretation of compounds in texts, provided that it is made more specific with domain-dependent or text-specific information.

2 Domain-independent model

In this section, we briefly explain how the interpretation is carried out when compounds contain explicit predicative information. We then focus on the interpretation of compounds in which the constituents are root nominals.

In what follows, semantic features are adapted from the WordNet lexical database[3] which provides a rich but non-specialized semantic taxonomy. We use a small part of this hierarchy in order to define a set of semantic features that label nominal constituents. Semantic labels are also used to express selectional restrictions on arguments.

2.1 Compounds with a deverbal constituent

Compounds including a deverbal constituent that subcategorizes the other constituent have been precisely described, in particular within the generative framework (Selkirk, 1982; Lieber, 1983). These results have been integrated in our model.

The predicatice relation between the constituents is given by the verbal root of the deverbal noun. We differentiate two types of deverbals: a deverbal may refer to the accomplishment or the result of the process denoted by the verb (e.g. *parsing*) or it may saturate the role assigned to one of the arguments of the verb and thus refer to one of the actors of the process (mainly agent or instrument, e.g. *parser*). In the former case (*action deverbals*), the deverbal inherits the entire argument structure of the verb; in the latter (*subject deverbals*), it inherits the structure minus the agent saturated by the suffix. When the deverbal noun occupies the head position of the compound, the non-head may saturate one of the roles of the argument structure of the deverbal, either the theme role, as in *sentence parsing* → *parse*(theme: *sentence*[4]), or a semantic role (in the sense of Selkirk (1982)), referring to a circumstance of the action (location, time, means, etc.): *hand parsing* → *parse*(means: *hand*). When the deverbal noun is the non-head, it cannot saturate an internal argument within the compound (Lieber, 1983); in this case, the head may only fill a semantic or an external argument: *parsing program* → *parse*(instrument: *program*).

This first series of compounding patterns has often be considered as the only type of compound which can be described in semantic terms (Selkirk, 1982). Our own position is to argue that the same predicate-argument pattern may be used to deal with other types of compounds, provided that we rely on a richer semantic representation of nominals, when no morpho-syntactic clues are available to constrain the semantic interpretation.

2.2 Root compounds

Nominal compounds illustrate the distributional properties of nouns in the absence of any explicit verbal predicate. They attest an underlying event structure associated to nominal constituents, which makes it possible to derive a predicative relation from the mere collocation of two simple nouns. The idea that noun meaning involves event-based description has been particularly emphasized by J. Pustejovsky (1991). We propose to apply a crucial component of his *generative lexicon*, the *qualia structure*, to the semantic interpretation of compounds.

The key idea that underlies the *qualia structure* is that nouns are implicitly related to predicative information, and that a noun selects for the type of predicate that can govern it. The four typical nominal relations that constitute the *qualia structure* are 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, that distinguishes the object within a larger domain.

We illustrate the use of this theoretical framework for the interpretation of nominal compounds.

Telic role. The notion of telic role is directly applicable to the treatment of compounds. It recalls Finin's notion of role nominals (Finin, 1980). A role nominal is typically linked to a verbal predicate that denotes its purpose; it fills one of the roles included in the argument structure of the verb. For example, the noun *pipeline* typically refers to the external argument of the verb *trans-*

[3] WordNet is a trademark of Princeton University.

[4] The semantic interpretation is represented in a formula that exhibits both the underlying predicate and the roles that each constituent plays in the argument structure of that predicate: N1 N2 → V(role_*i*: N2, role_*j*: N1). The head constituent is underlined.

port (cf. WordNet textual gloss: "a long pipe used to *transport* liquids or gases"). Unlike subject deverbals, role nominals are not provided with an argument structure that may be syntactically satisfied. Nevertheless, the argument structure of the underlying verb provides a clue for the distributional properties of the noun within compounds. The verb *transport* requires a subject and an object argument; since the noun *pipeline* refers to its first argument, the position which is left empty (the theme) may be occupied by the first constituent of a compound of the form *N pipeline*, as in *oil pipeline* → *transport*(instrument: *pipeline*, theme: *oil*).

Agentive role. The agentive role is also selected by the compounding mechanism: the non-head may refer to the origin of the head noun, as in *pancreas ptyalin* → *produce*(agent: *pancreas*, theme: *ptyalin*), in *compiler message* → *emit*(agent: *compiler*, theme: *message*), or in *bullet wound* → *cause*(agent: *bullet*, theme: *wound*). We see that this relation covers different kinds of predicates which are instances of a more general relation of creation.

Constitutive role. The constitutive role includes various kinds of semantic associations, such as part-whole relations (*outrigger canoe*) or substance relations (*stone house*).

Formal role. The formal role involves a relation of characterization which concerns different aspects of an object (its size, shape, color, etc.). The nouns that denote such information are mostly elements of the ATTRIBUTE class, which is defined in WordNet as "an abstraction belonging to or characteristic of an entity". Each member of this class may appear at the head position of compounds in which the non-head denotes the entity that is characterized: *desk height* → *characterize*(attribute: *height*, entity: *desk*). These nouns are uni-relational nouns that can appear as the head of "N1 of N2" groups, where N2 is a syntactic argument of N1 (e.g. *height of the desk*) (Isabelle, 1984).

Consequently, Pustejovsky's notion of noun's qualia helps to characterize implicit predicative link in compounds. This semantic framework demonstrates that the association between nominal constituents and underlying predicative relation in root compounds is not arbitrary: it involves conceptual mechanisms that are triggered in other linguistic phenomena such as type coercion (Pustejovsky, 1991), anaphora (Fradin, 1984) or adjectival constructions (Bouillon and Viegas, 1993).

2.3 Implementation and results

The implementation of these principles in our model is based on a conceptual framework in order to associate predicative information with nominal constituents. Two cases arise: when the link between a noun and a predicate is characteristic of a single noun, it is expressed in its lexical entry. When it is shared by a whole class of nouns, it is seen as a characteristic feature of that class which accounts for a relational property that any member of the class inherits. For example, the telic role of the word *pipeline*, which involves the verb *transport*, cannot be generalized to a whole class of nouns. On the contrary, the predicate CONTAIN is a characteristic feature of the class CONTAINER. Consequently, several predicates and several roles are potentially associated with nominal constituents, either as instances of different attributes, or as a consequence of this inheritance mechanism.

We have tested our model on a list of 100 compounds randomly picked up from a list of N N sequences in isolation[5]. Our program generates any interpretation that can be calculated on account of the mechanisms that we have described. Firstly, the list of predicates that are associated to the head constituent[6] is retrieved. Secondly, only the predicates that can provide a role to the other constituent are retained.

It is difficult to assess the correction of the answers that are produced, since we are dealing with compounds in isolation. Other answers are sometimes conceivable, if we apply less regular principles of semantic associations (Downing 1977), so that we cannot compare our results with a closed set of correct answers. Moreover, we cannot set a clear-cut border line between probable and hardly conceivable interpretations. Having said this, we can estimate our results as follows: 71% of the compounds that we have examined receive acceptable answers. For example, our program generates two clearly acceptable solutions for the compound *missile range*:

1) *characterize*(agent: *range_7*, theme: *missile*)
2) *shoot*(locative: *range_9*, theme: *missile*)

Contrary to Finin's and Mac Donald's models,

[5] This list of 9000 binary nominals has been kindly put at our disposal by R. Sproat. The corpus is described in (Sproat, 1994).

[6] In most cases, the predicative information is associated with the head, except when the non-head is deverbal, as in *hunting lodge*, or when the head refers to an underspecified event structure, as in *malaria program* (*fight*) vs *crop program* (*develop*). Such compounds illustrate the notion of co-compositionality (Pustejovsky 1991).

we are dealing with ambiguous constituents: nine meanings of the word *range* are listed, which correspond to the description given by WordNet for this noun. Only senses 7 ("scope", ATTRIBUTE) and 9 ("a place for shooting projectiles", ARTEFACT) are related to a predicative information that is compatible with the non-head, namely the formal role in the first case, and the telic role in the other. Some answers are more questionable:

cardboard box =

1) *constitute*(agent: *cardboard*, theme: *box_4*, *box_5*, *box_6*, *box_7*) - objects made of cardboard (constitutive role)

2) *contain*(locative: *box_7*, theme: *cardboard*) - box that contains cardboard (telic role)

3) *produce*(agent: *box_3*, theme: *cardboard*) - plant that produce cardboard (telic role)

4) *measure*(agent: *box_2*, theme: *cardboard*) - a quantity of cardboard (formal role)

Interpretations 2, 3 and 4 are surely mistaken in a standard context, if we refer to extralinguistic knowledge (*box_3* - a kind of shrub - does not produce cardboard the way *gum trees* produce gum) or to lexicalization (the compound *cardboard box* has only one usual meaning, namely *constitute*(agent: *cardboard*, theme: *box_7*, where *box_7* refers to the container). Yet, each answer is conceivable because it corresponds to productive semantic patterns and therefore to existing cognitive strategies.

6% of the answers miss expected answers and 23% give no answers at all. If we compare our results with those of Mac Donald (1982), we see that the part of silence is undoubtedly less important in his system (no meaning is produced for 10 % of the compounds). Nevertheless, one crucial distinction must be emphasized: in Mac Donald's system, slots are defined in relation to nominals, and an interpretation is identified if one constituent can fill a slot of the other. These slots are supposed to represent any piece of real-world knowledge that is necessary to understand noun compounds, but nothing precise is said about the information that needs to be stored. The solution to improve this result is unclear in such a system: missing interpretations correspond to absent slots, but no indication is given regarding the slots that must be added. On the contrary, we have shown that a few general principles of predicative attachment to nominal constituents are involved in the interpretation of compounds in our model; consequently, the analysis of incorrect answers allow us to determine in what cases domain-independent mechanisms are unsufficient to perform the interpretation and what kind of knowledge must be added to improve these results, either from domain-dependent or from contextual information. One can classify the problems in two categories:

Inappropriate selectional restrictions

Only selectional features can constrain the interpretation when several predicates are possible, in order to distinguish between different roles (e.g. *shoulder wound* - the non-head *affects* a BODY PART vs *bullet wound* - the wound is caused by a WEAPON). Consequently, no interpretation is generated when the semantics of the non-head does not match the constraints on the arguments of the predicate, and particularly in case of semantic shifts: *stadium* is a CONSTRUCTION, but in *stadium clash*, it is viewed as a LOCATION or as a GROUP of people. This is a general issue in lexical semantics; yet, the problem is all the more difficult to handle in compounds as no syntactic clue (i.e. no prepositional link) is available to distinguish between different (semantic or thematic) roles. It is also particularly problematic to solve ambiguous role assignment when semantic roles are concerned (as in *fear voters*).

Missing predicative link A general model cannot account for all possible compounding relations. Not to mention contextual links (Downing, 1977), some productive relations cannot be constrained from the semantics of the constituents. Specific links such as ressemblance (*carpet shark*) or subclass relations (*marathon tour*) cannot be described with structural rules. Moreover, a predicative information may be missed when it entails fine extralinguistic knowledge (e.g. *fruit fly*: insect whose larvae feed on fruits).

Generation of multiple interpretations and unpredicted patterns due to selectional violation or extralinguistic information are thus the two inherent limits of a domain-independent model of interpretation. Our aim is to give suggestions about the possibilities of refining this model when domain-specific or contextual information are available.

3 Domain-specific semantic information

3.1 Detection of specific patterns

Preferential patterns Statistical methods have been experimented by psycholinguists such as Pamela Downing (Downing, 1977) and Mary Ellen Ryder (Ryder, 1984): their purpose is to use statistical knowledge to interpret new compounds. Ryder argues that a set of semantic rules is not sufficient to deal with the productivity of the compounding process, since the creation of new compounds involves extralinguistic knowledge and cognitive strategies. According to her, "the predictability is probabilistic", and she shows that the creation and interpretation of new compounds is based on knowledge about productive semantic patterns. For example, she lists highly frequent templates such as:

N + PRODUCT = PRODUCT used on N (*pet shampoo, laundry detergent*)

This pattern illustrates only one facet - the telic one - of the head noun (and is irrelevant for examples such as *egg shampoo* or *dishwasher detergent*). This statistical result may differ considerably from one corpus to another. Consequently, frequency scores cannot be part of a domain-independent model.

From our results, we see that two types of specific information must be available to refine our domain-independent rules: firstly, we must specify the relative frequence of each role to assess the best interpretation for a compound when several semantic relations apply. Secondly, we want to determine the semantic features that characterize the non-head for one given role; P.Resnik's aim is similar when he illustrates the use of selectional association in compounds (Resnik 1993), in order to find N N semantic patterns which help to perform adequate bracketing of sequences with three constituents or more. He shows that it is difficult to find clear-cut semantic groups in unrestricted texts. Yet, such techniques, that combine statistic measures and conceptual knowledge, are very promising to exhibit typical patterns of association in specific domains.

Unpredicted patterns Exhibiting unpredicted patterns is a first step towards the determination of specific interpretation schemes in a given domain. For example, let us consider a list of compounds matching the N *pump* pattern, such as: *air pump, beer pump, breast pump, cattle pump, gear pump, piston pump, sand pump, stomach pump, drainage pump.* In this list, we find compounds exhibiting:

- the telic role of the noun:

SUBSTANCE + *pump* → *pump*(instrument: *pump*, theme: SUBSTANCE) (*sand, air*)

ACTION + *pump* → ACTION(instrument: *pump*) (*drainage*)

- the constitutive role of the noun

OBJECT + *pump* → *constitute*(theme: *pump*, agent: OBJECT) (*gear, piston*)

These patterns are predicted and interpreted by our set of rules. Other types of associations, too specific to be taken into account by our model, appear in the list: ANIMAL + *pump* (*cattle pump*) and ORGAN + pump (*stomach pump, breast pump*), in which the missing predicates are respectively *feed* - i.e. *pump food for* - and *clean* - i.e. *pump the contents of.* We see that the underlying telic relation is more complex, because it includes also an implicit argument (*food, contents*) of the predicate. These are typically the specific patterns that cannot be taken into account in a general model. Exhibiting semantic patterns in the texts is thus a way to automatically learn more specific patterns of associations in sublanguages. We are currently experimenting the way techniques of computer-aided acquisition for learning conceptual relations from syntactic collocates (Velardi et al. 1991) can be applied to N N associations.

3.2 Identification of the predicative link

Our model associates a fixed verbal predicate with nouns or nominal classes to account for a given semantic facet. This predicate corresponds to the typical predicative information that occur in the Wordnet textual gloss, when it is available. In fact, this predicate may vary from one corpus to another, and we must take into account this variation which corresponds to specific conceptual descriptions. Contextual information can contribute to identify the predicative relation by looking elsewhere in the text to see if the constituents of the compound are involved in another kind of linguistic construction, where their semantic relation would be explicit. Given a compound N1 N2, we may look for strings in which the couple (N1, N2) occurs in a different relation. In the following examples, the context provides the missing verbal predicate:

compiler warnings: (compiler,warning) = "it is reasonable for the compiler to *emit* a warning"

In this example, which corresponds to the agentive role, we see that the two nouns are arguments of the predicate that instantiates the underlying relation, which means that corpus-based methods can use a rich linguistic structure to identify the predicate. Pustejovsky et al. (1993) show how statistical techniques, such as mutual information measures can contribute to automatically acquire lexical information regarding the link between a noun and a predicate. Similar techniques are used by (Grefenstette and Teufel 1995) to determine the support verb associated with deverbal nouns.

Conclusion

This paper describes a domain-independent model for the interpretation of nominal compounds; it shows how general knowledge and domain-specific information may be combined for the interpretation of nominal compounds. Our goal is to account for productive and accross-domain rules of interpretation. Experimentation shows that the definition of general rules, which include conceptual description of the nominal constituents, implies the generation of multiple interpretations, especially since we are dealing with ambiguous nominal constituents.

We have proposed several ways of incorporating specific semantic information in our model, and we have suggested how corpus observations can detect preferential semantic relations and unpredicted semantic patterns. Statistical observations can contribute to identify the most productive compounding strategies for a given corpus, and are especially very promising as a way to deal with technical texts, in which the semantic variety of compounding relation is limited. This work is currently experimented in French, where it appears that the same conceptual framework holds to account for the semantic role of prepositions *à* and *de* in binominal sequences.

References

Pierrette Bouillon and Evelyne Viegas. 1993. Semantic Lexicons: the Cornerstone for Lexical Choice in Natural Language Generation, *Proc. of the seventh International Workshop of Natural Language Generation.*

Pamela Downing. 1977. On the Creation and Use of English Compound Nouns. *Language*, 53(4): 810-842.

Cécile Fabre and Pascale Sébillot. 1994. Interprétation sémantique des composés nominaux anglais et français, *Proc. of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition*, Genève.

Timothy Wilking Finin. 1980. The Semantic Interpretation of Nominal Compounds, *Proc. of the first conference of AI.*

Bernard Fradin. 1984. Anaphorisation et stéréotypes nominaux, *Lingua*, North-Holland, 64: 325-369.

Gregory Grefenstette and Simone Teufel. 1995. Corpus-based method for automatic identification of support verbs for nominalizations, *Proc. of EACL*, Dublin.

Pierre Isabelle. 1984. Another Look at Nominal Compounds, *Proc. of Coling-84.*

Rochelle Lieber. 1983. Argument Linking and Compounds in English, *Linguistic Inquiry*, 14(2): 251-285.

David B. Mac Donald. 1982, *Understanding Compounds Nouns*, PhD Thesis, Carnegie Mellon University.

Elaine Marsh. 1984. A Computational Analysis of Complex Noun Phrases in Navy Messages, *Proc. of Coling-84.*

James Pustejovsky. 1991. The Generative Lexicon, *Computational Linguistics*, 17(4): 408-441.

James Pustejovsky, Peter Anick and Sabine Bergler. 1993. Lexical Semantic Techniques for Corpus Analyses. *Computational Linguistics*, 19(2).

Philip Stuart Resnik. 1993. *Selection and Information: a Class-Based Approach to Lexical Relationships.* PhD Thesis, University of Pennsylvania.

Mary Ellen Ryder. 1994. *Ordered Chaos: the Interpretation of English Noun-Noun Compounds.* University of California Press.

Elisabeth Selkirk. 1982. *The Syntax of Words*, MIT Press.

Richard Sproat. 1994. English Noun-Phrase Accent Prediction for Text-to-Speech, *Computer Speech and Language*, 8: 79-94.

Wilco Ter Stal. 1996. *Automated Interpretation of Nominal Compounds in a Technical Domain*, PhD Thesis. University of Twente, the Netherlands.

Paola Velardi, Michela Fasolo and Maria Teresa Pazienza. 1991. How to Encode Semantic Knowledge: A Method for Meaning Representation and Computer-Aided Acquisition, *Computational Linguistics*, 17(2): 153-170.