知識工学と人工知能 56 － 4
（１９８８．１．１３）

概念に対する親しみの度合いを用いた意味理解

David Cohn*，藤澤浩道，木内伊都子

（株）日立製作所 中央研究所

従来のデータベースシステムなどの情報処理システムでは、複雑な概念を入力するためには、概念と概念の間の関係を陽に記述する必要がある。自然言語における複合名詞句、あるいは名詞節では、名詞(概念)相互間の関係の記述が少なくとも部分的に省略されている。従って、意味を理解するためにはその関係を推測する必要がある。本論文では、システムに記憶されている概念ネットワークで表現される知識を利用して、上記の陽に記述されていない概念間の関係を推論する新しい手法について述べる。さらに、本手法を用いて試作した、複合名詞句を含む名詞節から成る単純な自然言語文章を解釈する能力を持つ知識ベースインターフェースについて述べる。

The Use of "Familiarity" in Semantic Interpretation

David COHN*, Hiromichi FUJISAWA, Itsuko KIUCHI
Central Research Laboratory, Hitachi, Ltd.
1-280 Kokubunji, Tokyo 185, JAPAN

Abstract

Most information handling systems require that the relationships between concepts be explicitly stated. Nominal compounds and noun phrases are natural language structures in which the relationships between concepts are at least partially omitted and must be inferred. We describe a knowledge base interface capable of semantically interpreting simple sentences comprised of nominal compounds and short noun phrases based on the system's familiarity with the relations between the component concepts.

---

*currently on leave from University of Washington, Seattle, WA 98195, USA

1. Introduction

1.1 Basic Problems in Filing

User interfaces tend toward two extremes in their adaptation toward the user and the machine he is using. Formula and frame-based systems, while simple for the machine to process, are a major obstacle for the novice or casual user. The other extreme is the full-blown natural language understanding (NLU) system which, while intending to simplify the user's task, enormously complicates the machine's side of the problem, frequently resulting in large plodding programs that take up more space than the original system. To make matters worse, these complete systems generally still lack the ability to let a novice user communicate in truly natural language without a good understanding of the knowledge base structure, and frequently get him tied up in syntactic formulae as complex as those that are needed by a non-NLU interface.

A possible middle road is presented here: have the user avoid complicated syntax by communicating in short sentences made up of nominal compounds and noun phrases. These short, succinct chunks of information are naturally constructed and used by people in everyday conversation. Some examples are "Find articles on personal computer software packages" and "The 2050 is a new Hitachi workstation."

When a nominal compound is used, the relationship between its component parts has been omitted and must be inferred by the hearer, in this case, the computer system. Noun phrases, while providing some clues to the nature of their missing relations, still require considerable inference to interpret correctly. Although there are no existing theories that describe how to reliably infer these missing relations and their grouping, the problem appears to be amenable to being mapped onto a concept-relation network.

1.2 Overview of the Approach

The target knowledge base for this interface was the UNIFILE system [Fujisawa], a hierarchical concept-relation network with multiple inheritance. It was designed as an intelligent interface to an optical disk filing system. The network allows users to browse through general information as well as knowledge gleaned from summaries of stored articles. The articles themselves exist as concepts in the network and can be retrieved for viewing in digitized image form. At present, the full network contains some 372 registered articles and almost 2500 concepts.

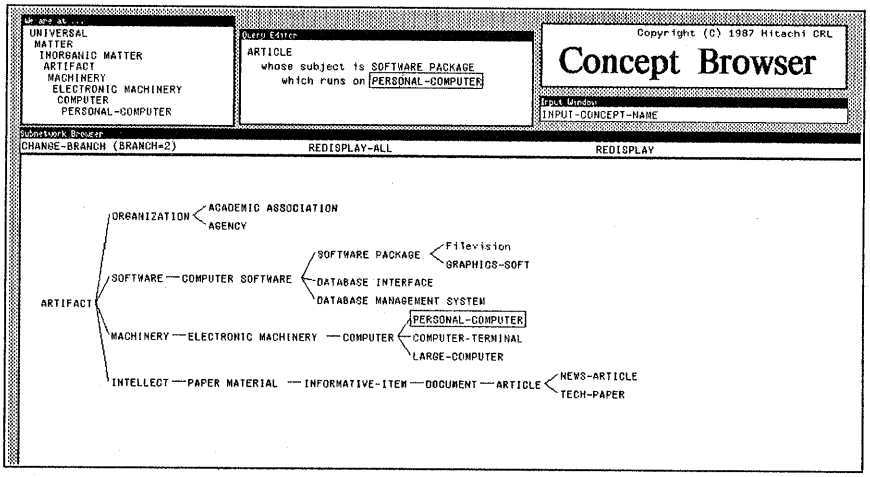

fig 1: A sample screen from a session with UNIFILE

Concepts in the network, which may have any number of aliases, are connected by hierarchical is-a and is-part-of links into a partial ordering. Non-hierarchical generic relation links connect abstract concepts in the network to each other via links that may be traversed in either direction. The concepts "person" and "published-material," for example, are connected by the "authorship" relation which is read "is-author-of" in one direction and "is-written-by" in the other. These relations are inherited by child concepts, and concrete concepts at the bottom of the hierarchy (such as specific people and books), are connected by instantiations of these relations. Input and queries are managed through a "concept browser" which allows searches to be organised into frames of information centered around relevant concepts.

The DLink interface (for "Dynamic Linking") is a limited natural language understanding system supported by a simple chart parser based on [Thompson]. It allows interpretation of sentences constructed from simple noun compounds and noun phrases. Interpretation is primarily semantic with a minimum amount of syntactic constraint.

Section 2 of this paper describes the "principle of familiarity," a metric used by the DLink system to aid in the interpretation of input. Section 3 describes how the principle is used to understand nominal

compounds, and Section 4 shows how it is applied to understand complete noun phrases. Section 5 describes the idea of structured concepts as implemented by DLink. Excerpts from an actual session and conclusions are presented in Sections 6 and 7, respectively.

2. Familiarity as a Metric

The use of system "familiarity" in interpretation is suggested by the observation that in practice, available knowledge acts as a context that people use as a tool for constructing and interpreting noun phrases. Since the knowledge available to a system is limited to its data base, this leads to the hypothesis that the system can use its own knowledge about concepts and their relations as an interpretation tool. This leads to a sort of "system pragmatics," akin to the pragmatics that humans use to fill in missing or ambiguous information. Of course, in the machine, this interpretation must be accomplished using knowledge of a greatly reduced scope. The world of the system is its list of concepts and relations, and by basing its interpretation on this sub-world, it is seeking a "practical interpretation."

The principle of familiarity, as defined here, states that the strength of a particular link between two concepts is measured as a function of the number of instantiations of that link or similar links in the knowledge base. While it may at first seem to be an arbitrary case of "aiming at the largest target," this appears to be a good heuristic for modelling the interpretation methods of people.

For example, most people would interpret "American computer company" to mean an American company that deals with computers. While the idea of company that produces American computers is quite conceivable, the idea of an "American computer" is not as strongly instantiated in the mind as that of an "American company" and a "computer company."

3. Nominal Compound Interpretation

In this application, the standard definition of a "nominal compound" has been extended slightly to mean any string of nouns, adjectives, and nominalized verbs with no intervening particles. "Computer company" is a simple example; "January automobile water pump cover shipments" [Tennant] is considerably more complex.

Nominal compound interpretation can be broken into two tasks: determining the implicit links between pairs of words and "bracketing" the compound into consistent word pairs. To illustrate, deriving the meaning of "Japanese computer company" requires deciding upon the relation between "computer" and "company" (as in "computer is produced by company"), and deciding whether the compound should be bracketed as "a company that produces Japanese computers" or as "a computer company that is Japanese."

Much research in both computational linguistics and computer science has focused on the interpretation of nominal compounds. Early attempts at "handling" such compounds [McDonald], amounted to paraphrasing the compound at a lexical level, providing no real insight as to its implicit meaning. While some recent efforts at natural language understanding attempt to accommodate actual interpretation [Rich], most natural language interfaces appear to specifically exclude their use.

The problem is still an open area of research, but several tendencies in the formation of such compounds allow a heuristic approach to be taken in their interpretation. The links normally omitted in a compound can usually be classified into one of several categories that can be checked by a semantic network. Bracketing, in addition to having a left-associative tendency, is restricted semantically, as will be discussed later in this section under "Nominal Compound Bracketing."

3.1 Determining Implicit Links

3.1.1 Generic Links

Research on interpreting implicit links suggests that a majority of compounds tend to be formed when the concepts are generically linked, or "when the relationship in question is of habitual nature" [Downing]. In the UNIFILE knowledge base, concepts are defined in terms of what relations they have to one another as regulated by a hierarchical set of generic relations. This representation scheme lends itself quite well to the testing of Downing's observations. Since the concept "article" is defined in terms of such relations as "has subject X" and "is part of journal Y," these generic links to X and Y can be applied when hearing about a "supercomputer article" or an "ElectronicsWeek article."

On the basis that synonyms are common in real life, the UNIFILE system does not require that concept names be unique. When synonyms are encountered, possible links to both concepts are investigated. The query weighting is based on the overall best fit, so even if a spurious partial match has a very high weight, its inability to fit the entire compound smoothly will detract from its selection. If the query building process reaches a dead-end, it abandons that structure and continues processing on the compound's other possible interpretations.

For example, according to the DLink system, the term "computer articles" has three possible interpretations: articles whose subject is computers, articles that are part of Computer magazine, and articles that mention computers. Counting examples of each interpretation, however, the system decides that it is most familiar with articles whose subject is computers, and lists that as its first choice.

3.1.2 Generalized Links

From the list of possible generic links between concepts, the system chooses only those that have actually been instantiated for some instances of the two concepts. This is useful when a compound is being used referentially, i.e., to refer to a pre-existing concept. Frequently, however, a compound will be used in an introductory sense, where the concept or the relevant links may not yet exist.

When referring to "the 5 MHZ CPU" referentially, it is assumed that the hearer already knows about a number of CPUs, one of which is known to run at 5 MHZ. When a compound is used to introduce a concept, some inferences must be made. For example, when the system is first told about image processors and mention is made of a "10 MHZ image processor," there are no links connecting any image processors to the concept "10 MHZ." Either a new link of some sort must be established between an existing image processor concept and the relevant cycle speed, or a new image processor concept must be created that is compatible with existing links.

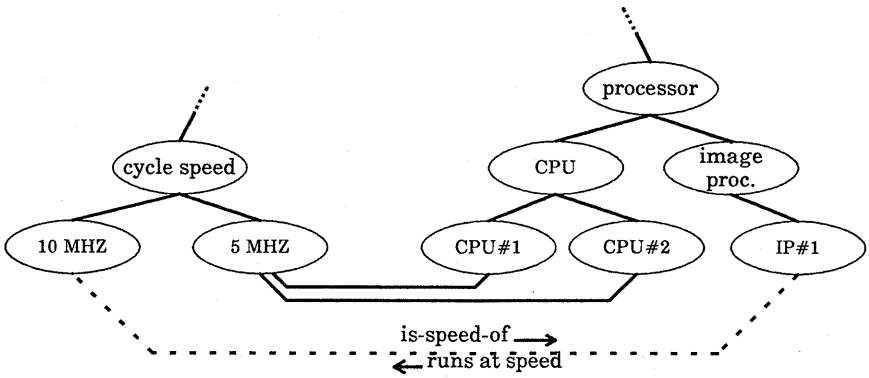

fig 2: Generalizing an analogous relationship to infer a new link

The approach taken here is one of finding examples by successive generalization. When it is determined that there are no existing links between image processors and "10 MHZ," the next levels up in the hierarchy, "processor" and "cycle speed" are checked. Here, all instances of links between processors and cycle speeds are examined. Since the concept "CPU" is also a child of "processor" and the system has previously created links connecting CPUs to cycle speeds, it can make the inference that the new compound is being used in the same way.

3.1.3 Non-Generic Links

It is important to note that while generic links form the basis for most nominal compounds, some compounds are based on clearly non-generic relations. These compounds are usually non-lexicalized, novel compounds that are created because of some specific event which relates the two components.

An example which Downing gives is that of a friend who, having once parked her bike in the building's vestibule, was thereafter referred to as the "bike girl." This was not a habitual, nor an expected relation between "bike" and "girl." Still, because of a specific event, the relation assumed by "bike girl" was considered sufficient for identification of the person in question.

Such non-generic relations are managed in UNIFILE by searching "event nodes," special nodes created to describe complex events. These nodes center around the action being done, parking for example, and are linked to other concepts in terms of function they serve in the event: the actor, object, time, place and any other action specific information that the node might require.

If a particular computer had been purchased in South Dakota, for example, then a reference to the "South Dakota computer" is meaningful, even though there are no direct links connecting the concepts "South Dakota" and "computer." The purchase itself, registered as an event concept, has links to the computer and the place the purchase took place, as well as any other links considered relevant to the event.

3.2 Nominal Compound Bracketing

In a simple, two element compound, interpretation involves merely selecting the most promising link between the two concepts. Many compounds, however, consist of three or even many more elements, so the problem of correct bracketing the internal associations must be handled.

In English, a word in a compound modifies one of the words it precedes, meaning that all words eventually modify the final word of the compound. This last word, called the "head noun" by linguists, generally determines the "type" of the whole compound. For example, a newspaper article is an "article," and a Japanese personal computer company is a "company." It should be noted that this is an assumption made in

the name of simplicity, and exceptions do arise. These exceptions seem to be cases where the head noun has become lexicalized, such as "hot dog", or assimilated into the compound itself, as in "a Renoir" (painting). These cases can usually be covered by creating lexicalized compounds as aliases, or by defining them as structured concepts , as described in Section 5.

(((Japanese personal) computer) company) :	Strict left bracketing; "Japanese personal" doesn't have any possible interpretation.
(Japanese (personal (computer company))) :	Strict right bracketing; "personal" modifying "company" is unlikely.
((Japanese (personal computer)) company) :	A company producing Japanese personal computers.
(Japanese ((personal computer) company)) :	A Japanese company producing personal computers.

fig 3: Possible bracketings of "Japanese personal computer company"

As seen in figure 3, the lack of an appropriate relationship (such as between "Japanese" and "personal"), will eliminate a number of candidate linkings. The structured, parenthetical representation solves the "no crossing of branches" rule for ruling out impossible interpretations. This rule prevents any association where "personal" modifies "company" and "Japanese" modifies "computer". Such an interpretation, crossing modifiying branches, would create an improperly nested structure, and thus will not be created by the bracketing routine.

Using the above methods to find candidate links, we will be left with a number of possible links that can be made. Starting with the head concept as the root, a nested structure is built up sequentially, starting from the end of the compound. The left-bracketing tendency in English can be emulated when building backwards by trying to associate new concepts at the most deeply nested level in the structure. The no crossing of branches rule is enfored by only considering the last concept at any level for deeper association.

For example, figure 4 shows part of the placement sequence for the compound "new American LISP computer company." The concept "computer" is associated with "company" by the "production" relationship. When the concept "LISP" is to be attached, it is first tried at the deepest level, associating it with "computer". This succeeds, but when "American" is associated with the new deepest level, "LISP," the match is very weak. Trying successively larger scopes, the system settles on "American company," a concept which is well instantiated in the network. (The concept "American" is defined both as the concept of an American [person] and as the adjectival form of the concept "America." In this case, the latter definition was selected by the system.) It is important to note that the next placement, the word "new," is tried first on "America" and then on "company." Although "LISP" is actually a more deeply nested concept, because "computer" is no longer the last concept at the second level, it and all of its modifying concepts are excluded by the no crossing of branches rule.

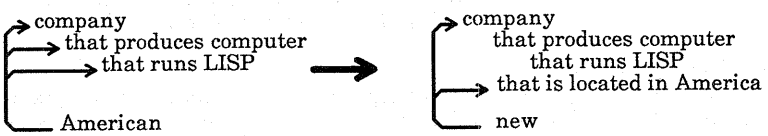

fig 4: Successive modifier attachment rules out anomalous interpretations

4. Noun Phrase Interpretation

One problem with interpreting nominal compounds is that their internal structure is not delimited, and the number of possible interpretations that need to be checked tends to rise exponentially. Beyond compounds several words long, unambiguous understanding becomes difficult for both humans and computers. For this reason, humans usually break longer compounds into several pieces (called "noun groups," here), connected by simple modifiers. "American personal computer software package articles" would generally become the noun phrase "articles *on* software packages *for* American personal computers" (delimiting modifiers italicized).

It is instructive to note that in actual human conversation, a novel idea is often presented first in the form of a noun phrase, such as "articles about software," to instantiate it in the hearer's mind. After this, the compound "software articles," can be used in the conversation as a matter of course.

4.1 Determining Noun Phrase Links

Noun phrase interpretation requires much the same approach as nominal compound interpretation with the exception that the modifiers between concepts in the phrase serve as clues to what kind of relation is intended. These modifiers are associated with the noun or nominal compound directly following them, and knowing the type of this concept can significantly narrow the choice of relations suggested by the modifier. In the phrase "the book by Newton," the modifier "by" could suggest any one of a number of relations, but the fact that it is attached to the concept "Newton," a person, leaves "authorship" as the only possible relation.

This authorship relation must have, as its other end concept, some form of written material, and this knowledge allows sets of plausible links to be constructed as with the nominal compound case, only with the greater reliability provided by the "clues."

The function served by modifiers in joining noun phrases can actually be served by a broader class of words that we have labeled "joiners." Joiners in the DLink system are by no means restricted to single words; they can even become complete verbal clauses. The only restriction is that they be a string of words that semantically imply the relationships that they call up. Some examples of joiners that call up the authorship relation are "by," "from," "was written by," and "whose author is ." The correspondence between relations and joiners is many-to-many: a single relation may be suggested by a number of joiners, and a simple joining phase (such as "from"), may suggest many different relations.

The actual number of joiners needed to provide reasonable coverage in recognizing a particular relation is kept down by generalizing routines that neutralize tense and number differences in phrases for matching purposes. If a joiner is not recognized by the system, the user can register this new phrase with the system by following prompts.

4.2 Noun Phrase Bracketing

In contrast to nominal compounds, the head, or root of a noun phrase is the leftmost noun group. Additional modifying joiners and their associated noun groups are added on to the structure formed by this group from left to right, following the same placement strategy as is used for nominal compounds. In the classic ambiguous statement "I saw the man on the hill with the telescope" this strategy automatically eliminates the anomalous interpretation that plagues semantic parsers.

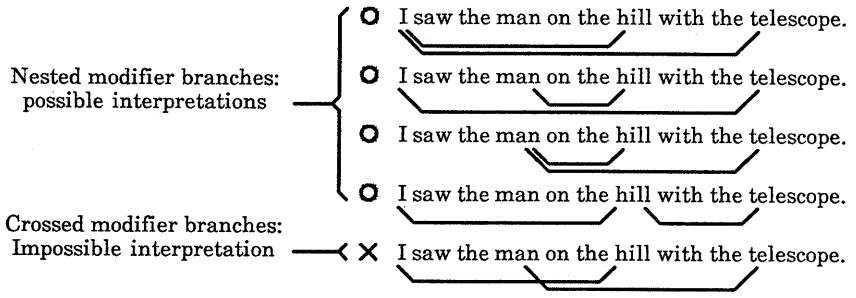

fig 5: Building a structured representation
avoids the "no crossing of branches" problem

4.3 Noun Phrase Parsing

The parsing strategy is aimed avoiding as much as possible having a grammar independent of the knowledge base. All the aliases for referring to concepts are stored in the network, as are all the joiner phrases. For example, in the generic relation definition table [Fujisawa], the "authorship" link has all joiners that suggest authorship aliased as names of its directional links. Additional function words, such as articles and quantifiers, exist as independent concepts in the network with all necessary semantic information attached directly to them.

relationship	left-right	right-left
authorship	is -author-of	author-is
	is-the-author-of	by
	wrote	from
	has-written	of

fig 5: Aliases attached to the directional links
of a generic relation.

Another feature that allows the system to operate with a minimal grammar is the assumption that its input will be correct and relatively unambiguous. By limiting input to short sentences and noun phrases, most of the need for syntactic differentiation of tenses and plurals can effectively be ignored. The parsing routine is syntactic in that it does not discard unrecogizeable input, but the range of syntactic errors it can accommodate is rather large.

The separate grammar that is needed is thus very lenient in terms of syntax, relying on semantic information to weed out improper interpretations later. This allows the actual parsing to be done with a very simple chart-style parser based on [Thompson].

5. Structured Concepts

As mentioned earlier, some nominal compounds have an actual type that differs from the type of their head noun. One example would be a reference to a "Renoir," an implicit compound that actually refers to a "painting by Renoir." This form of compound appears to be idiosyncratic, with no reliable way of infering the missing concepts, but the problem can be handled on a case by case basis using "structured concepts."

A structured concept (roughly equivalent to Brachman's "defined concepts"), is one that is defined as a template built up of other concepts and relations between them. This structure is stored separately from the concepts and relations it refers to and becomes part of the separate grammar used by the interface. Once a structured concept for a "Renoir" is created by defining it as a "painting which was painted by Renoir," all paintings in the network that have the necessary link to the French artist will be considered by this reference.

In addition to assisting with idiosyncratic nominal compounds, structured concepts have a variety of useful applications in the interface and in the UNIFILE system as a whole. The hierarchical structure of the system, even with multiple superclasses, makes representation of some concepts difficult. The concept "author" could be defined as a profession, with appropriate links defining it as the job position of people who have written books, or those people could be attached as children of the "author" node. Much more intuitively appealing, however, is the creation of the structured concept "author" defined as "a person who has written a book," which allows all people who fit this requirement to be automatically included by the reference to "author." Structured concepts can be simple, as in "a potential-customer is a person," or arbitrarily complex, as in "a potential-customer is a person at a company located in the USA who owns a personal computer that runs LISP."

The structures themselves, existing separately from the concept network, can also be used to personalize one's view of the knowledge base by defining complex ideas relevant to one's own interests. This allows the single, central system to accommodate a variety of very different users, each with their own preferred nomenclature and abbreviations.

6. Sample Results

The following are excerpts from a session using the DLink interface to interact with the UNIFILE system. Some system maintenance dialog has been left out for brevity. User input is underlined and comments are shown as side notes and inline enclosed in square brackets.

```
[* User creates a new concept *]
DLink > "Robert Lowe" is an American man.
"Robert Lowe"
Have parsed an instance frame for:   -- note: DLink asks for confirmation before creating
man                                           any new concepts or links.
    whose nationality is U.S.A.;
Instantiate it? y

[* User adds new information to an existing concept *]
DLink > Robert Lowe is also known as "Bob".
DLink > Bob is a professor at the University of Washington.
Have parsed an instance frame for:
"Robert Lowe"
   whose job position is professor;
   which works at University of Washington;
Instantiate it? y
professor        Assistant Professor -- note: The concept "Professor" has a number of possible
                 Associate Professor          interpretations, so the system asks for clarification.
                 Donner Professor
                 Professor

 Robert Lowe whose job position is professor: 1
 Robert Lowe whose job position is Assistant Professor: ok

[* User defines a structured concept *]
DLink > Define author as a person who has written a book.
DLink > Find all authors at American companies.
--There are 85 candidate concepts under person.
Q: person
    which is author of book;          -- note: Structured concepts, like author, maintain their
    which works at company            structure independent of the network.
       which is located in U.S.A.;
```

--6 concepts match the above abstract concept.

[* User inputs an ambiguous request *]
DLink > Find articles on American personal computer software packages.
article
  whose subject is software package
    which runs on a personal computer;
    which was developed at organization.workplace
      which is located in U.S.A.;
is one of 2 possible interpretations.
Is this one acceptable? y

--There are 68 candidate concepts under article.
Q: article
    whose subject is software package
      which runs on personal computer;
      which is developed at organization.workplace
        which is located in U.S.A.;

--2 concepts match the above abstract concept.
[* At this point, the user can issue a system command to retrieve the digitized article images from the optical disk subsystem for viewing. *]

7. Conclusions

The primary purpose of this research is to provide enhanced user-computer interaction capabilities; the method by which this is being pursued is through the development of a nominal compound and noun phrase interpreter capable of limited natural language understanding. The concept of familiarity as a heuristic for interpreting missing or ambiguous semantic links is proposed here and has been implemented as part of this interface.

The results of preliminary testing indicate that nominal compounds and noun phrases form a useful language subset that human users can easily formulate and computers can interpret with reasonable accuracy. The familiarity heuristic is a valuable part of the system and supports the hypothesis of "system pragmatics" as a method of emulating human interpretation.

Structured concepts are a simple constructs that, in addition to aiding the natural language interface, provide a powerful and useful tool for knowledge representation. They allow personalizing the knowledge base and aid in manipulation of complex concepts that can not easily be accommodated in a strict hierarchy.

As a result of the implementation strategy, it has been found feasible to support natural language parsing from within the structure of a knowledge base, and avoid having to support a large external language grammar.

Bibliography

[Brachman] Brachman, R.; Schmolze, J., "An Overview of the KL-ONE Knowledge Representation System," Cognitive Science, Vol. 9, No. 2, 1985, pp. 171-216

[Downing] Downing, P., "On the Creation and Use of English Compound Nouns," Language, Vol. 53, No. 4, 1977, pp. 810-842

[Fujisawa] Fujisawa, H.; Hatakeyama, A.; Higashino, J., "A Personal Universal Filing System Based on the Concept-Relation Model," Proc of the 1st Int. Conference on Expert Database Systems, Charleston, SC, 1986, pp. 31-44

[Isabelle] Isabelle, P., "Another Look at Nominal Compounds," Proc. of Coling84, Stanford, CA, 1984, pp. 509-516

[McDonald] McDonald, D.; Hayes-Roth, F., "Inferential Searches of Knowledge Networks as an Approach to Extensible Language-Understanding Systems," in Pattern-Directed Inference Systems, edited by D. Waterman and F. Hayes-Roth, Academic Press, New York, NY, 1978, pp. 431-453

[Rich] Rich, E; Wittenburg, K.; Barnett, J.; Wroblewski, D., "Ambiguity Procrastination," Proc. of AAAI, Seattle, WA, 1987, pp. 571-576

[Tennant] Tennant, H., Natural Language Processing. Petrocelli Books, New York, NY, 1984

[Thompson] Thompson, H.; Ritchie, G., "Implementing Natural Language Parsers," in Artificial Intelligence: Tools, Techniques, and Applications, edited by T. O'Shea and M. Eisenstat, Harper and Row, New York, NY, 1984, pp. 245-300