DICTIONARY TEXT ENTRIES AS A SOURCE OF KNOWLEDGE FOR SYNTACTIC AND OTHER DISAMBIGUATIONS

Karen Jensen and Jean-Louis Binot[1]

IBM Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, New York 10598

Abstract

Online reference books may be thought of as knowledge bases. We describe here how information in the text of machine-readable dictionary entries can be processed to help determine the proper attachment of prepositional phrases and relative clauses; the resolution of some cases of pronoun reference; and the interpretation of dangling modifiers. This approach also suggests the possibility of bypassing conventional efforts at hand-coding semantic information, efforts which are time-consuming and usually incomplete.

0. INTRODUCTION

Online reference books may be thought of as knowledge bases, with data structures encoded in natural language. We have developed a system that reasons heuristically about the comparative likelihood of various potential attachments for prepositional phrases in English sentences by analyzing relevant definitions in Webster's online dictionary (W7) in their original text form (Binot and Jensen 1987, Jensen and Binot forthcoming). This paper reviews that earlier work and then extends it by suggesting how additional information (particularly example sentences from another dictionary, the Longman Dictionary of Contemporary English (LDOCE)) might be used to cope with three additional problems: the attachment of relative clauses, the resolution of some cases of pronoun reference, and the interpretation of dangling modifiers. The earlier work on PP attachments has been implemented, but we have only begun work on the implementation of these additional disambiguation problems. Nevertheless, it seems like a good idea to indicate that this dictionary-based approach should be feasible for more than PP attachments.

Our objective is to consult the dictionary to find the kind of information that has previously been supplied by means of scripts, frames, templates, and other hand-crafted devices. This approach offers hope for reducing time-consuming, and usually incomplete, hand-codings of semantic information; and it should be of particular interest for non-restricted text processing applications such as machine translation and critiquing.

We are concerned here with emulating, in some sense, the way a person uses a dictionary: look up one entry, study the definitions and the examples, look up other entries, and so on. We feel that natural language itself can be a reasonable knowledge representation language. More needs to be learned about how to access and manipulate this knowledge; but the flexibility afforded by natural language is an advantage for the task, not a drawback.

This research is related to other work being done with machine-readable dictionaries, e.g. Markowitz et al. 1986, in the sense that we all share the goal of automatically extracting semantic information from these rich sources. However, in other respects our approaches are quite different.

1. ATTACHMENT OF PREPOSITIONAL PHRASES

The relationships in which we are interested can be illustrated by the following sentences from Binot 1985:

(1) I ate a fish with a fork.
(2) I ate a fish with bones.

(See Appendix A, Tree 1.) In both cases, the ambiguity resides in the placement of the "with" prepositional phrase, which might modify either "fish" or "ate". The parse tree shows the PP attached to the closest possible head, "fish," with a question mark showing that it could alternatively be attached to the verb "ate".

[1] The second author currently works for B.I.M., Belgium.

Focussing on (1), another way to phrase the key question is "Is it *more likely* that a fork is associated with a fish or with an act of eating?" To answer that question, the system evaluates separately the plausibility of the two proposed constructs:

(1a) eat with a fork
(1b) a fish with a fork

then orders the solutions, and picks the one with the highest rating.

In the heuristics we are currently using, the basic way to rate the likelihood of a construct is to try to establish, through the dictionary, some relevant semantic connection between the words of that construct. Easier (or shorter) connections yield better ratings. Long connections, or connections making use of approximate inferences, will lead to lower ratings. For example, the definition of "fork" contains the phrase "used for taking up," and "eating" is defined as a kind of "taking" in the dictionary. By establishing these relationships, we see a plausible semantic connection between "fork" and "eat," and (1a) receives a high rating.

The relationships are established (a) by identifying equivalent function-word patterns in the definitions, such as the equivalence of "used for" and the instrumental "with"; (b) by linking important definition words (i.e., central terms in definitional phrases, such as heads of phrases, or else synonyms). This is done by parsing the definitions, identifying the central word(s), and then following hierarchical chains of definitions through the dictionary.

Heuristic answers are expressed in terms of *certainty factors* which, as in the MYCIN system (Shortliffe 1976), take their values in the range (-1, +1): "-1" expresses absolute disbelief; "0" expresses complete uncertainty; "1" expresses absolute belief. Intermediate values express varying degrees of belief or disbelief.

The two main heuristics that are used to evaluate the plausability of (1a) against (1b) can be described in English as follows:

H1- for checking for an INSTRUMENT relation between a head and a "with" complement:

1. if the head is not a verb, the relation doesn't hold (certainty factor = -1);
2. if some "instrument pattern" (see below) exists in the dictionary definition of the complement, and if this pattern points to a defining term that can be linked with the head, then the relation probably holds (certainty factor = 0.7);
3. else assume that there is more chance that the relation doesn't hold (certainty factor = -0.3).

H2- for checking for a PARTOF relation between a head and a "with" complement:

1. if the head is not a noun, the relation doesn't hold (certainty factor = -1);
2. if some "part-of pattern" (see below) exists in the dictionary definition of the complement, and if this pattern points to a defining term that can be linked with the head, then the PARTOF relation probably holds (certainty factor = 0.7);
3. else assume that there is more chance that the relation doesn't hold (certainty factor = -0.3).

Each certainty factor refers to the specific proposition (or goal) to which the heuristic is applied. Thus, if clause 3 of heuristic H2 is used when applied to the proposition (1b), the resulting certainty factor -0.3 will indicate a relatively moderate disbelief in this proposition, stemming from the fact that the system has not been able to find any positive evidence in the dictionary to sustain it.

The above heuristics make use of the fact that there are specific words and/or phrases in dictionary definitions, forming *patterns*, which are almost systematically used to express specific semantic relations (Markowitz et al. 1986). For the two relations considered here, some of these patterns are:

INSTRUMENT: for, used for, used to, a means for, etc.
PARTOF: part of, arises from, end of, member of, etc.

These patterns generally take, as their objects, some central term (or terms) in the definition of the complement word. We can then try to link that term with the head of the construct that is being studied.

Focussing again on example sentence (1), the system starts by examining the first construct,

(1a). It parses the definition of the complement "fork," and discovers at least one INSTRUMENT pattern, "used for":

fork: An implement with two or more prongs *used esp for taking up* (as in eating), *pitching* or *digging*.

Taking the conjunction into account, the system finds three possible terms: "taking up," "pitching," and "digging," which it tries to link with "eat." (For the present, we deliberately avoid the phrase "as in eating" -- which offers a direct match -- in order to show that our approach does not rely on such lucky coincidences.) The system is able to establish that "eat" is a direct hyponym of "take" according to W7:

eat: to *take* in through the mouth as food... to *take* food or a meal.

The link is thus probably established, and the system moves on to consider (1b). Since no PARTOF pattern can be found in the definitions of "fork," this second possible construct will be ranked as much less likely -- (1a) receives a certainty factor of +0.7, but (1b) gets a certainty factor of only -0.3. Therefore the system recommends attaching the PP to the main verb in (1).

For sentence (2), the constructs to be compared are "eat with bones" and "a fish with bones." In the definition of "bone," no useful INSTRUMENT pattern is found; so "eat with bones" cannot be easily validated. But the first definition of "bone" gives the following PARTOF pattern:

bone: One of the hard *parts of the skeleton of a vertebrate*.

This yields two possible links for "fish": "skeleton" and "vertebrate." "Fish" can be identified as a direct hyponym of "vertebrate" according to W7.

fish: Any of numerous cold-blooded strictly aquatic craniate *vertebrates*...

Therefore, "a fish with bones" receives a higher certainty factor than "eat with bones," and the system recommends attaching the prepositional phrase to the direct object in sentence (2).

The above examples are among the simplest. In more difficult cases, heuristics may perform various kinds of inferences in order to establish connections. It is also possible for several heuristics to be applied to a given construct, with their results then being combined. The cumulative effect of many heuristics, and not the perfection of each one separately, does the job.

The choice of certainty factors rests mainly on intuition. Some choices are easy; some inferences, for example, are obviously weaker than others. In other cases the values have to be adjusted by trial and error, by processing many examples. It is interesting to note that, as our corpus of examples increases, the certainty factors are converging toward apparently stable values. Our system currently includes about 20 heuristic rules and is able to handle the prepositions "with," "by," "after," and "in." It has been tested successfully on about 50 examples so far.

2. ATTACHMENT OF RELATIVE CLAUSES

A typical problem in attaching relative clauses occurs when the clause is separated from the noun it modifies by a prepositional phrase:

(3) I want the book by my uncle that is on the shelf.

In (3), the relative clause "that is on the shelf" probably modifies "book" and not "uncle." A human reader assumes this because of knowing that a book is more likely to be on a shelf than an uncle is. However, syntax alone cannot tell us so. A syntactic parser will normally produce a tree which shows the relative clause modifying the closest noun, namely "uncle."[2] (See Appendix A, Tree 2.) Note that the parser attaches the relative clause (RELCL) node arbitrarily to the closest head noun "uncle," but marks the other possible attachment site ("book") with a question mark. The higher question mark in Tree 2 is for the PP attachment.

[2] The grammar that supports all of the parsing discussed here is the PLNLP English Grammar (Jensen in preparation, Heidorn 1976).

We have implemented the solution to this kind of relative clause ambiguity. Our system starts by trying to solve the PP attachment problem: does "by my uncle" modify "book" or "want"? Of all possible relationships between the various word pairs, the AUTHOR relationship between "book" and "uncle" will receive by far the best ranking. This will happen because it can be established, by using the dictionary, that an uncle can be a human being (and thus able to author a work), and that a book is some kind of work.

The processing of the RELCL attachment then begins. Syntax tells us that the relative pronoun "that" is the subject of the predicate "be on the shelf." One of the properties of the verb "to be" is that a prepositional complement qualifying this verb really qualifies the subject of the verb. Applied to Tree 2 of Appendix A, this provides two possible interpretations:

book on the shelf
uncle on the shelf

At this point we can see that the relative clause attachment in Tree 2 reduces to a prepositional phrase attachment, which can be solved easily by the PP attachment methods already described. Specifically, the dictionary definition for "shelf" will tell us that a shelf is "to hold objects" or "for placing things on," and the word "book" can be related to "object" or "thing" much more easily than the word "uncle" can be so related. This will lead to the preference for "book" as the antecedent of the relative clause.

However, most relative clause attachment problems cannot be reduced to PP attachments. Consider (4):

(4) I know the actor in the movie that you met last month.

The parse tree for this sentence (Tree 3 of Appendix A) shows question marks in the same positions as Tree 2. However, because of the syntactic structure of the RELCL in (4), we know that the relative pronoun this time refers to the object of its main verb "met." Either "movie" or "actor" must be the object of "met." No prepositional phrase is involved.

Now we have to decide which is more likely:

You met an actor.

You met a movie.

Although semantic codes are included in the online version of LDOCE (i.e., features like HUMAN and ABSTRACT are marked on nouns, and subcategorization codes using these features are marked on verbs), the codes do not help with problems like this one. According to the LDOCE codes, possible objects for the simple transitive verb "meet," in its various sub-senses, are HUMAN, ABSTRACT, and (moveable) SOLID. No ranking of likelihood or preference is given, and of course a syntactic parser would not know which sub-sense it is dealing with. "Actor" is marked +HUMAN, and "movie" is marked +ABSTRACT. So either object noun is equally likely (Mary Neff, personal communication).

Although we have not yet implemented this, we believe that the same "approximate reasoning" that we implemented for PP-attachments will work here, too. The strategy is to formulate heuristics that yield "certainty factors," not categorial acceptance or rejection of an interpretation. These heuristics would propose a solution for the stated task by operating on the output of the syntactic parser. For the current example, the first step would be to parse the LDOCE entry for "meet" (shown in Figure 1), looking for direct objects.

meet[1] /mi:t/ *v* **met** /met/ **1** [T1;I0] to come together (with), by chance or arrangement: *Let's meet for dinner.*|*I met him in the street* —compare **MEET WITH** **2** [T1] to find or experience; **MEET WITH**: *I met a lot of difficulties in the work* **3** [I0] to come together or close: *The cars almost met* **HEAD-ON** (=one front against the other), *but drew away and drove on* **4** [T1,I0] to get to know or be introduced (to) for the first time: *Come to the party and meet some interesting people.*|*We met at Ann's party, didn't we, but I don't remember your name* **5** [I0] to join at a fastening point: *My skirt won't meet round my middle* **6** [I0] to gather together: *The whole school met to hear the speech* **7** [T1;I0] to touch, (as if) naturally: *Their lips met* (*in a kiss*).|*Her hand met his face in a violent blow* **8** [T1 (*with*)] to answer, esp. in opposition: *His charges were met with cries of anger.*|*Angry cries met his speech* **9** [T1] to be there at the arrival of: *I'll meet you off the train.*|*The taxi will meet the train/will meet you off the train* **10** [T1] to pay: *Can you meet this amount?* **11** [T1] to satisfy: *Does this meet your hopes?*|*This new road meets a long-felt need* **12 make ends meet** to use one's (small amount of) money carefully so as to afford what one needs **13 meet someone's eye** also **look someone in the eye**—to look directly or steadily at someone **14 more** (**in/to something**) **than meets the eye** hidden facts or reasons (in or for something) —see also **meet HALFWAY**

Figure 1. Text of LDOCE entry for the verb "meet"

The sub-definitions are no help, because no objects are shown. But the example sentences in the entry are a rich source of information about typical usage. There are eleven different example object nouns: *him, lot (of difficulties), people, face, speech, you* (twice), *train, amount, hopes, need.* Over a third of them can be easily related to the word "actor": the word "people," and the three occurrences of personal pronouns. (The general rule here is that any personal pronoun except "it" can be substituted for any word that has "person" as the head of one of its definitions.) None of them can be so easily related to the word "movie." Thus the system concludes that "actor" is a more likely object of the verb "meet" than is "movie." This conclusion is no accident; lexicographers are experts on words, and they have incorporated their expertise, in ways both obvious and subtle, into standard dictionaries.

Another interesting example of the relative clause attachment problem is found in the following sentence from a large data base of business letters:

(5) There are no agencies within the country which would loan money to individuals for establishment of boarding homes.

The choice here is between possible nouns to serve as the subject of the predicate "would loan money":

Agencies would loan money.
Country would loan money.

First, the LDOCE definition for "loan" refers us to the word "lend." Moving to the entry for "lend," we look for cited subjects. The example sentences, in this case, are no help: subject words are either personal pronouns or the word "flags"; and none of these helps us to choose between "agencies" and "country." But one of the sub-definitions of "lend" is

"to give out (money) for profit, esp. as a *business*".

The phrase "as a" is often used in definitions to signal the AGENT that does the action. Then we consult the dictionary to see which better qualifies as a "business": "agency" or "country." The answer comes easily; the first sub-definition of "agency" is

"a *business* that makes its money esp...".

The two words "country" and "business" cannot be connected so easily as "agency" and "business" along any path of heuristic searching. Therefore we prefer to attach the relative clause to "agencies" rather than to "country."

It is important to realize that none of the information being cited here is manually coded; the English text of the LDOCE entries is being used. Our strategy can be considered to be making explicit a semantic network that exists implicitly in this text. The entry for "lend" shows "business" as an AGENT of "lending"; the entry for "agency" shows that "agency" is a kind of (ISA) "business." This implicit chunk of network is shown in Figure 2:

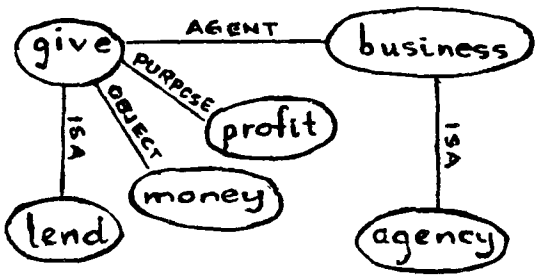

Figure 2. A semantic path connecting "lend" and "agency" in LDOCE

3. RESOLUTION OF PRONOUN REFERENCE

Problems of pronoun reference are many and varied, and not all of them will yield to this same method of solution (Hobbs 1986, Sidner 1986). But for some, the information in dictionary definitions can give important clues. Consider (6) and (7):

(6) We bought the boys apples because they were hungry.
(7) We bought the boys apples because they were cheap.

In the absence of other information, human readers assume that "they" probably refers to the boys in (6) and to the apples in (7). The computer needs to follow some inference path that will lead to the same tentative assumptions.

For sentence (6), we need to choose a most likely subject noun for the predicate "be hungry" -- either:

Boys were hungry.
Apples were hungry.

We would first parse the dictionary definition for "hungry." In LDOCE, there are two example sentences with personal pronouns for subjects; the word "boys" can be quickly related to all personal pronouns. There are no example sentences with subjects that can be easily related to "apples."

Additional support can be found in two directions. The first definition for "hungry" in LDOCE is "feeling or showing hunger." We want to find out what sort of entity can "be hungry," so we ask what sort of entity can "feel." Of about 30 example sentences for the verb "feel," 26 are personal pronouns (excluding "it"). Hence we prefer "boys" to "apples" as the subject of "be hungry."

A second direction of search also reinforces this interpretation. "Hungry" is defined as "feeling or showing hunger," and "hunger" is defined as "the wish or need for food." Briefly summarized, we conclude that "food" is the object (or goal) of hunger, hence of being hungry. LDOCE also tells us that an "apple" is "a hard round fruit" and "fruit" is "used for food." Hence apples are (used for) food; hence apples can be the object of "being hungry." Since the suggested object of "being hungry" is the same as the object of the main clause (see (6)), it stands to reason that "they" probably does not also refer to "apples."

The paths that we are tracing are delicate, but they exist. A computer program that follows these paths extracts, from existing text, some very interesting real-world relationships.

In solving the pronoun reference task of sentence (7), the program must choose between:

Boys were cheap.
Apples were cheap.

By following paths through the LDOCE entries, the conclusion that "apples were cheap" appears more likely than that "boys were cheap" (although the latter is certainly possible).

4. INTERPRETATION OF DANGLING MODIFIERS

English teachers have long objected to a potential awkwardness and lack of clarity in constructions with dangling modifiers:

(8) (While) watching TV, the doorbell rang.

In sentences like (8), the attachment problem appears in a different guise. There is only one noun given for the participial to modify, and that is "doorbell." It is not possible to set up an obvious choice pair in the same manner as before. However, we do know that participial modifiers are a notorious source of confusion. So we can check the dictionary to find out how likely it is that a doorbell might watch TV.

In LDOCE, the sub-definitions for "watch" are no help. But the example sentences, once again, offer strong hints. There are 16 such examples. Fifteen of them have personal pronouns as subjects for the verb "watch." The first example is "Do you often watch TV?" (This situation was not contrived; sentence (6) was taken from a popular high school English grammar book, Warriner 1963, before the dictionary was consulted.) With this information in hand, we can say that "doorbell" is, at best, an unlikely subject for the verb "watch."

5. CONCLUSIONS

There are many important sources of information for natural language processing. Syntax, logical form, intersentential context, and presuppositions about the mental state of the speaker and of the intended audience (to name a few) all make their contributions, and have all been discussed, to varying extents, in the literature. Now it appears that the text portions of online dictionary entries can serve as a rich source of semantic information and world knowledge that can aid during the processing of other text.

ACKNOWLEDGMENTS

We would like to thank George Heidorn for his helpful suggestions in the preparation of this paper, and Yael Ravin for her continuing insights into problems of ambiguity.

REFERENCES

Binot, Jean-Louis. 1985. *SABA: vers un systeme portable d'analyse du francais ecrit.* Ph.D. dissertation, University of Liege, Liege, Belgium.

Binot, Jean-Louis and Karen Jensen. 1987. "A semantic expert using an online standard dictionary." *Proceedings of IJCAI-87*, Milan, Italy, August 1987.

Heidorn, George E. 1975. "Augmented phrase structure grammars" in Nash-Webber and Schank, eds., *Theoretical Issues in Natural Language Processing*. Association for Computational Linguistics.

Hobbs, Jerry R. 1986. "Resolving Pronoun Reference" in Grosz et al., eds., *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, Inc., Los Altos, CA.

Jensen, Karen. In preparation. "PEG: A broad-coverage computational syntax of English." IBM Research Report.

Jensen, Karen and Jean-Louis Binot. forthcoming. "Disambiguating prepositional phrase attachments by using on-line dictionary definitions." *Computational Linguistics*, special issue on the lexicon.

Longman Dictionary of Contemporary English. 1978. Longman Group Ltd., England.

Markowitz, Judith, Thomas Ahlswede, and Martha Evens. 1986. "Semantically significant patterns in dictionary definitions." *Proceedings of the 24th Annual Meeting of the ACL*, Columbia University, June 1986.

Shortliffe, E.H. 1976. *Computer-based medical consultation: MYCIN.* Artificial Intelligence Series. Elsevier.

Sidner, Candace L. 1986. "Focusing in the Comprehension of Definite Anaphora" in Grosz et al., eds., *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, Inc., Los Altos, CA.

Warriner, John E. 1963. *English Grammar and Composition: Complete Course.* Harcourt, Brace & World, Inc., New York.

Webster's Seventh New Collegiate Dictionary. 1963. G. & C. Merriam Co., Springfield, Mass.

Appendix A. Parse Trees

```
DECL   NP      PRON*   "I"
       VERB*   "ate"
       NP      DET     ADJ*    "a"
               NOUN*   "fish"
        ?      PP      PREP    "with"
                       DET     ADJ*    "a"
                       NOUN*   "fork"
       PUNC    "."
```

Tree 1. Parse tree for a syntactically ambiguous PP attachment

```
DECL   NP      PRON*   "I"
       VERB*   "want"
       NP      DET     ADJ*    "the"
               NOUN*   "book"
        ?      PP      PREP    "by"
                       DET     ADJ*    "my"
                       NOUN*   "uncle"
                ?      RELCL   NP      PRON*   "that"
                               VERB*   "is"
                               PP      PREP    "on"
                                       DET     ADJ*    "the"
                                       NOUN*   "shelf"
       PUNC    "."
```

Tree 2. Syntactic parse showing relative clause attachment

```
DECL   NP      PRON*   "I"
       VERB*   "know"
       NP      DET     ADJ*    "the"
               NOUN*   "actor"
        ?      PP      PREP    "in"
                       DET     ADJ*    "the"
                       NOUN*   "movie"
                ?      RELCL   NP      PRON*   "that"
                               NP      PRON*   "you"
                               VERB*   "met"
                               NP      AJP     ADJ*    "last"
                                       NOUN*   "month"
       PUNC    "."
```

Tree 3. Another relative clause attachment