Representing Information Need with Semantic Relations*

Anil S. Chakravarthy

MIT Media Laboratory

Abstract

Information retrieval systems can be made more effective by providing more expressive query languages for users to specify their information need. This paper argues that this can be achieved through the use of semantic relations as query primitives, and describes a new technique for extracting semantic relations from an online dictionary. In contrast to existing research, this technique involves the *composition* of basic semantic relations, a process akin to constrained spreading activation in semantic networks. The proposed technique is evaluated in the context of extracting semantic relations that are relevant for retrieval from a corpus of pictures.

1 Introduction

In recent years, there has been considerable interest in applying techniques from computational linguistics to improve various aspects of information retrieval (IR) [1]. This paper describes new techniques for extracting semantic knowledge from online linguistic resources in order to provide better methods of expressing IR queries.

Belkin and Croft [2] used the term *information need* to characterize the user's motivation for using an IR system. In current IR systems, users first translate their information need into queries. The IR system processes these queries and matches them against surrogates representing the text (or media) collections to retrieve elements of the collection that are possibly the best matches to the user's queries.

In most information retrieval systems in wide use, the surrogates for the collection are words or word collocations that are either extracted automatically or provided by a human indexer. This constrains users to express their queries as combinations of words or word collocations, resulting in an inaccurate or inadequate description of their information need. Sneaton [1] characterizes the ideal as a conceptual information

retrieval system, wherein users express their need as some combination of concepts and the system matches these to concepts representing the underlying text or media collection. Attempts have been made to build such systems for specific domains, but how to automatically extract and represent concepts in general is still far from clear.

Many researchers in computational linguistics have recognized that electronic dictionaries could be used to address this conceptual information bottleneck (e.g., [3, 4]) and a lot of work has been devoted lately to extracting semantic relations between words (e.g., [5]). This research allows queries to be expressed as combinations not only of words but also of semantic relations. To take a simple example, let us assume that the system has access to a database of pictures of animals with the names of animals as the surrogates. If a user is interested in pictures of dogs, the query [X A-KIND-OF dog] is constructed. The system processes the query to retrieve pictures whose surrogates are words like "basset hound," "beagle," and so on since the query matches the semantic relations, [basset hound A-KIND-OF dog], [beagle A-KIND-OF dog], etc. Thus, semantic relations enable the user to construct queries that correspond to entire classes of word-based queries.

The main purpose of this paper is to describe new techniques for extracting semantic relations that were inspired by the work of Quillian [6]. Quillian demonstrated that by organizing individual semantic relations into semantic networks, one could obtain compositions of existing semantic relations by a process of spreading activation. For example, the two relations, [basset hound A-KIND-OF dog] and [dog HAS-PART tail], can be composed to yield [basset hound HAS-PART tail].

In this paper, we describe a program in which individual semantic relations extracted from a dictionary are composed to yield new semantic relations for retrieval over a database of pictures. We address three issues that arise in this connection:

- Control of spreading activation: Unbounded spreading activation often results in connecting words through relations that do not fall into any desired type. To be useful, it is necessary to control spreading activation so that only relations of

the desired types are found.

- Equivalence of alternative compositions: Depending on the configuration of the semantic network, there might be several acceptable alternatives that yield the same new composed relation.

- Word-sense ambiguity: Since individual semantic relations are extracted from the dictionary text, it is necessary to constrain the spreading activation to the "correct" senses of the target word.

Section 2 describes the test database that we have been using and semantic relations that are useful for retrieval over this database. In Section 3, we describe a pattern-based approach that we have employed to control spreading activation and recognize alternative compositions. In Section 4, we present the results and analysis of a series of tests that we conducted to test the accuracy of the program and its coverage as compared to a hand-constructed system, WordNet [7]. This section also describes and evaluates a new word-sense disambiguation technique that is based on knowledge of the semantic relations involving the ambiguous word. Finally, Section 5 includes a brief summary of the work and discusses issues that need to be addressed in future work.

2 A Database of Pictures

The primary motivation for our work was to provide retrieval based on semantic relations for a corpus of pictures collected from the American Heritage Dictionary. The corpus contains 1359 pictures, each of which is annotated with a single word or word collocation from the dictionary. Clearly, there are a great many semantic relations that could be useful for retrieval from such a database. To narrow down the set of interesting semantic relations, we used the fact that the annotations are single words or word collocations. As in memory experiments in cognitive psychology, we used the annotations as cues for free recall by association. We then analyzed the results to locate semantic relations that occurred most often. Based on this analysis, we picked the seven relations shown in Table 1 (which we will henceforth call *modes* to distinguish them from individual semantic relations).

The OCCURS-WITH mode refers to typical physical collocation of objects. It is useful for making "intelligent" guesses about what else might be in the picture besides the objects explicitly annotated. As the example in Table 1 shows, this is not always symmetric. It can be argued that the presence of an ax in a picture much more often indicates the presence of wood than the other way around. The PLAYS-ROLE-OF mode differs from the EXAMPLE-OF mode in having a connotation of typical use. The CHARACTERISTIC-ACTIVITY mode is used to relate both objects and agents to typical activities they

Mode	Example
OCCURS-WITH	(ax, wood)
EXAMPLE-OF	(basset hound, dog)
PLAYS-ROLE-OF	(cat, pet)
CHARACTERISTIC-ACTIVITY	(ax, chopping)
HAS-PURPOSE	(aqueduct, water)
CONSTITUENT-OF	(balance beam, gymnastics)
HAS-CONSTITUENT	(dog, leg)

Table 1: Modes used for picture database retrieval

are involved in. The HAS-PURPOSE mode is used to relate an object to a word denoting its purpose. As in the Table 1 example, that word could either denote an activity or another object where there is a typical activity involving both objects. CONSTITUENT-OF and HAS-CONSTITUENT are similar to the widely-used PART-OF and HAS-PART primitives except that metaphorical inclusion is valid as well. The next section describes our scheme for extracting these modal relations from the dictionary[1].

3 Extracting Modal Relations from Dictionary Definitions

Extracting modal relations from dictionary definitions involves three components: a preprocessor that tags the definition with part-of-speech information, a module that pulls out *triples* (basic semantic relations of the form [word1 LINK-TYPE word2]) from the preprocessed definition, and a pattern interpreter that checks the list of triples for modal relations using sets of patterns. We will now describe each of these in turn.

For preprocessing the dictionary definitions, we have experimented with two different taggers: the Xerox PARC part-of-speech tagger [8], and the Chopper [9], an optimizing finite state machine-based tagger built at the MIT Media Lab by Ken Haase. Before tagging the definition, we apply a few simple filters to remove botanical names, usage guidelines, etc. The performance of both taggers was satisfactorily high. The example below shows the output of the Xerox tagger on a sample definition[2]:

```
aqueduct: a    conduit for water
         :AT :NN      :IN :NN
```

1. Use library of A-KIND-OF or ENTAILS extraction patterns to locate the genus term. Extract triples from modifiers of the genus term.

2. Iterate over the differentiae constructing triples using each of them until either the end of the definition or till no triple can be constructed from the differentia found.

3. Apply post-processing methods to construct other triples using matching rules.

Figure 1: Procedure for extracting triples from a pre-processed definition

3.1 Extracting Triples

The algorithm for extracting triples is built on the assumption that dictionary definitions[3] typically consist of a genus term (identifying its kind) followed by differentiae (how it is different from the genus) [10]. In the aqueduct definition above, the genus term is "conduit" ([aqueduct A-KIND-OF conduit]) and the only differentia is given by the PP "for water". In the case of verb definitions, the genus term is related by the link "ENTAILS". The three stages of the algorithm are presented in Figure 1.

We use a variety of patterns described in the literature to extract the initial genus term(s) correctly [5] (e.g., patterns like "a NP," "either of two plural-NP," "one of a family of plural-NP"). The patterns combine both syntactic and string elements, which makes them more powerful than purely string-based patterns [11]. Since it is very important to find the genus term correctly, a "last-ditch" extractor is invoked if none of the standard patterns work. This last-ditch extractor assumes that the tagger must have made a mistake and tries to compensate for common tagger mistakes (e.g., tagging an ing-form as verb instead of adjective). Once the genus terms have been found, we analyze the morphological form of the modifiers for triples. For instance, since "violin" is defined as "a bowed instrument," the triple [violin OBJECT-OF bowing] is recorded[4].

In Step 2, each of the differentiae is assumed to be either a relative clause or a prepositional phrase. As in Step 1, head nouns or verbs are located for each of the differentiae and result in triples being formed with the word(s) being modified. Where there is attachment ambiguity (as with prepositional phrases [13]), triples are formed for all possible attachments.

Step 3 is a post-processing step which results in new triples being formed and some triples from

Step 2 being eliminated. For instance, consider the following definition of "acropolis": "acropolis: the upper, fortified part of a Greek city (as Athens)." In Step 3, two triples produced in Step 2, [acropolis A-KIND-OF part] and [part OF city] are merged into [acropolis PART-OF city]. Similarly, there are rules for creating links of other types. Other post-processing rules deal with eliminating references to A-KIND-OF genus terms in triples by replacing them with the definiendum.

3.2 Extracting Modal Relations from Triples

For each modal relation in Table 1, there is a set of patterns that specifies how the modal relation can be detected from the triples of one or more definitions. Each pattern encodes a heuristic rule that is based on the triples extracted from the dictionary definitions. For example, one pattern for extracting OCCURS-WITH relations encodes the heuristic that typically if Object1 and Object2 are involved in the same action, [Object1 OCCURS-WITH Object2]. When this pattern is applied to the definition, "ax: a cutting tool that consists of a heavy edged head fixed to a handle with the edge parallel to the handle and that is used especially for felling trees and chopping and splitting wood", the two modal relations [ax OCCURS-WITH tree] and [ax OCCURS-WITH wood] are found. Patterns can apply to multiple definitions as well. Using the same heuristic as above, we have defined a pattern that extracts the modal relation [atomizer OCCURS-WITH spray] from the two definitions below:

> **atomizer**: an instrument for atomizing usually a perfume, disinfectant, or medicament
> **atomize**: to reduce to minute particles or to a fine spray

4 Performance Evaluation

To analyze the performance of the program, we picked the first 300 annotations from the corpus of pictures described in Section 2. The first part of this section presents the results of applying the modal relation patterns that work on one definition. The second part discusses difficulties involved in the evaluation of modal relation patterns that work on more than one definition.

In order to obtain an independent estimate of the performance of the extraction program[5], we compared the output of the program where possible to the output of WordNet [7], which is a large, manually-coded semantic network. WordNet does not have links corresponding to four

---

[3]The program only handles noun and verb definitions.

[4]In all, there are about 15 link types in triples, namely, A-KIND-OF, ENTAILS, PART-OF, HAS-PART, AGENT-OF, OBJECT-OF, WITH, FOR, AS, OF, and several spatial prepositions like IN and ON [12].

[5]i.e., independent of the dictionary we are using;

ext	cor	%	wnt	wnm	%	#p
471	321	68.15				8
870	742	85.28	537	138	25.69	4
12	9	75.00				2
212	129	60.84				6
146	114	78.08				6
114	75	65.78	99	5	0.05	2
126	83	65.87	124	4	0.03	4

Table 2: Performance of the modal relation extraction program

of the modes listed in Table 1: OCCURS-WITH, PLAYS-ROLE-OF, HAS-PURPOSE, and CHARACTERISTIC-ACTIVITY. For the other three modes, we found correspondences by assuming that all hypernyms are valid examples of EXAMPLE-OF, all meronyms of HAS-CONSTITUENT, and all holonyms of CONSTITUENT-OF. The performance results and the comparison with WordNet are presented in Table 2.

The seven rows of Table 2 correspond to the seven modal relations of Table 1. The first column shows the total number of modal relations extracted for a mode while the second column gives the number of modal relations judged to be correct (by the author) with the percentage figure shown in the third column. The fourth column gives the number of such relations found in WordNet, while the fifth gives the number of those relations that were also found by the extraction program (with the column after that providing the percentage figure). The last column shows the number of modal extraction patterns implemented for the mode.

We will now briefly discuss the performance of the program. A detailed analysis is presented in [15]. The precision of the extraction is over 60% in all cases; there are three main reasons for the precision not being higher:

- Many of the patterns (e.g., for OCCURS-WITH or CHARACTERISTIC-ACTIVITY) implicitly assume that verbs denote activity. This is not true of many verbs like "suggest", "represent", "resemble", etc.

- Many patterns hinge on the presence of particular links (like "WITH" and "IN"), and precision is dragged down by their ambiguity.

- The tagger makes mistakes during the preprocessing resulting in incorrect matches for the patterns.

The number of matches with WordNet was generally low because WordNet uses word collocations as link destinations to construct more detailed hierarchies. So, for instance, while WordNet has the link [accordion HYPERNYM free-reed instrument], our program generates [accordion EXAMPLE-OF instrument].

We have not similarly analyzed the performance of patterns that operate over two definitions. The main reason is that to get an accurate estimate of the precision (as in the second column of Table 2), we have to combine the dictionary definition of the test word with every other definition in the dictionary. This work is in progress. However, it is clear that word-sense ambiguity can lead to poor performance by running modal extraction patterns over unintended senses of a word. For instance, if we return to the "atomizer" example at the end of the previous section, we find that we are "spreading activation" through the verb "atomize." There is another sense of "atomize", viz., "to subject to atom bombing," which is not of interest here and should be ignored. We will now briefly describe a new word-sense disambiguation technique that is applicable in this context. A detailed discussion can be found in [15].

Krovetz and Croft [14] characterized the process of word-sense disambiguation as bringing to bear several kinds of evidence depending on the context of occurrence of the word, namely, part-of-speech, morphology, subcategorization, semantic restrictions and subject classifications. Continuing in the same framework, we decided to use the semantic relations involving an ambiguous word as another source of evidence. Let the ambiguous word be denoted by W_{amb}, and $[W_{src}$ RELATION $W_{amb}]$ be a triple in the first definition of the modal relation pattern. Then, each definition of W_{amb} which includes the triple $[W_{amb}$ RELATION-INVERSE $W_{src}]$ can be considered as a correct sense (for spreading activation), where RELATION-INVERSE is the inverse link type of RELATION[6]. For the same 300 words as in Table 2, we tested this hypothesis on three kinds of links:

1. A-KIND-OF: The inverse of A-KIND-OF is AS. The definitions of "building" and "structure" given below illustrate the inverse relationship.

   building: a usually roofed and walled structure built for permanent use (as for a dwelling)
   structure: something (as a building) that is constructed

2. PART-OF, whose inverse is HAS-PART.

3. HAS-PART, whose inverse is PART-OF.

The results were very disappointing, with less than 5% of the words tested being successfully disambiguated by this technique. Often the problem seemed that the inverse link was present, but using a synonym or a hyponym. To test this, we conducted an experiment on HAS-PART where all we required to judge a

---
[6] Clearly, this technique only applies to dictionaries and other text sources which are definitional in nature.

definition as correct was the presence of some PART-OF triple, no matter what it was part of. This made the technique too general and ineffective. Out of the 103 HAS-PART relations tested, there were 14 correct disambiguation cases and 35 incorrect cases.

A more effective technique seems to be to use the modifiers of the ambiguous word. We conducted an experiment on A-KIND-OF links in which, for an ambiguous definiendum, we accepted as the correct senses those definitions whose genus terms had some modifiers in common with the definiendum in its original context. Out of the 860 entries tested, there were 84 correct disambiguations and 39 incorrect ones.

5 Conclusions

This paper is based on the argument that semantic relations can provide a better way for users to express their information need to an IR system. For such an interface to be widely applicable, the IR system should be capable of automatically acquiring semantic relations from available linguistic resources rather than through hand coding. Recent work in semantic knowledge extraction from online dictionaries seems like a promising method of automatic acquisition. This paper describes a new technique for extraction from dictionaries that was inspired by the well-known mechanism of spreading activation in semantic networks. The technique relies on sets of patterns to compose basic semantic relations from one or more dictionary definitions into *modal relations* between words. We evaluate this technique in the context of seven modes that are useful for retrieval from a database of pictures. Finally, we propose and evaluate a new technique for word-sense disambiguation within dictionary definitions that makes use of semantic relations involving the ambiguous word.

We plan to extend this work in many directions. We would like to develop a standard grammar for defining modal relation patterns, and an interpreter for this standard grammar. Once the grammar is available, we intend to develop an interface for specifying new modal relations and patterns for detecting them over multiple definitions. We also plan to incorporate implementations of existing word-sense disambiguation algorithms into the modal pattern interpreter. From an IR perspective, we plan to investigate the synergistic use of modal relations with other query primitives like keywords for retrieval from an extended database of annotated pictures. One desirable result of this work for future research would be the establishment of standard benchmarks and test suites for retrieval from picture databases.

References

[1] Smeaton, A. F. "Progress in the Application of Natural Language Processing to Information Retrieval Tasks," in *Comp. Jnl.*, **35**(3), 1992.

[2] Belkin, N. J., and Croft, W. B. "Information Filtering and Information Retrieval: Two Sides of the Same Coin?," in *CACM*, **35**(12), 1992.

[3] Fox, E. A., Nutter, J. T., Ahlswede, T., Evens, M., and Markowitz, J. "Building a Large Thesaurus for Information Retrieval," in *Proc. of the Second Conference on Applied NLP*, 1988.

[4] Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, A., and Slator, B. "Providing Machine Tractable Dictionary Tools," in *Semantics and the Lexicon*, J. Pustejovsky, editor, Kluwer Academic Publishers, 1993.

[5] Chodorow, M. S., Byrd, R. J., and Heidorn, G.E. "Extracting Semantic Hierarchies from a Large On-Line Dictionary," in *Proc. of ACL-23*, 1985.

[6] Quillian, M .R. "Semantic Memory," in *Semantic Information Processing*, M. L. Minsky, ed., MIT Press, 1968.

[7] Miller, G. A., and Fellbaum, C. "Semantic Networks of English," in *Cognition*, **41**, 1991.

[8] Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P. "A Practical Part-of-Speech Tagger," in *Proc. of the 3rd Conf. on Applied NLP*, 1992.

[9] Haase, K. B. "Multi-Scale Parsing using Optimizing Finite-State Machines," Internal Memo, MIT Media Laboratory, 1993.

[10] Amsler, R. "The Structure of the Merriam Webster Pocket Dictionary," PhD Dissertation, University of Texas, Austin, 1980.

[11] Montemagni, S. and Vanderwende, L. "Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries," in *Proc. of COLING-92*, 1992.

[12] Miller, G. A., and Johnson-Laird, P. N. *Language and Perception*, Harvard U. Press, 1976.

[13] Jensen, K., and Binot, J-L. "Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions," in *Computational Linguistics*, **13**, 3-4, pp 251-260, 1987.

[14] Krovetz, R., and Croft, W. B. "Word Sense Disambiguation Using Machine-Readable Dictionaries," in *Proc. of SIGIR-89*, 1989.

[15] Chakravarthy, A. S. "Representing Information Need with Semantic Relations," Internal Memo, MIT Media Laboratory, 1993.