A Hybrid Approach to Restricted Text Interpretation

Marti A. Hearst

Computer Science Division, Evans Hall
University of California, Berkeley
Berkeley, CA 94720

Abstract

One way to extract meaning from a large text corpus is to interpret documents in terms of a restricted semantic model. This paper describes a hybrid approach that integrates the techniques of information retrieval with ideas from cognitive linguistics. In contrast to the standard text processing system, the goal of which is to discover documents that pertain to some *topic* of interest to the user, an approach is introduced based on the criterion of *directionality* (e.g. "Is the agent in favor of, neutral, or opposed to the event?"). A method is described for coercing sentence meanings, based on their syntax and content, into a metaphoric model that can be easily interpreted in order to answer a direction-based query. The assumptions and tradeoffs associated with this method are discussed.

Introduction

The increasing availability of computer-accessible text is intensifying the need for innovative ways to process this special kind of data. To avoid the expense of a full semantic analysis, one can *restrict* the type of information extracted from the text in the hopes of obtaining a savings in processing time and complexity. This is feasible if useful criteria for restricting the semantics can be developed, and if the bulk of the text processing is done by efficient, but coarse, techniques. This paper proposes such a hybrid method, integrating the ideas of information retrieval with those of cognitive linguistics. This description is meant to be more a vision for future viable systems than something that can be implemented in full today, mainly because it assumes the existence of successful implementations of components that are currently still research goals.

The proposed method, called Task-Based Text Interpretation (TTI), interprets isolated portions of larger texts within the framework of a general, domain independent metaphoric model. In fact, TTI uses metaphor in two ways: (1) as the basis for a simple model into which the semantics of sentences are mapped, and (2) as a criterion to determine how to place semantically related lexical items into groups, or conglomerations. These conglomerations circumvent the need for the large, complex knowledge bases that full text understanding systems require. Applying this method to the output of an Information Retrieval (IR) system yields an incremental improvement in the text classification task.

The remainder of the paper first motivates TTI with a description of the task domain, then gives a general overview of the method. This is followed by an exposition of the role of general metaphor in the system and an application of TTI to a sample sentence. Finally, since any semantic analysis done in the context of an IR system must strike a favorable balance between the amount of processing required and the depth of interpretation obtained, the paper concludes with a discussion of the assumptions and tradeoffs found in this approach.

Motivation and Problem Statement

One way of structuring a corpus is to sort the documents into categories based on their topical content. Current IR systems accomplish this task with varying degrees of sophistication. For example, RUBRIC (McCune *et al.* 1985) allows the user to define an elaborate conceptual hierarchy, bottoming out on keywords, that classifies documents according to what topics they contain. However, once the topical content of a document is determined, how can the document be further distinguished from others describing the same topic?

What is needed is a classification criterion that applies to a wide range of corpora; a useful question rel-

atively independent of domain. One such criterion is: where, according to the text, does a semantic attribute lie along a continuum between extremes? For example, given a set of newspaper articles and the topic "Environmental Issues Pertaining to the Kesterson Wildlife Refuge," one can inquire as to whether public figures are stated as being opposed to, neutral, or in favor of a proposed cleanup plan. More generally, articles can be classified according to how they answer the query "Is agent A in favor of event E?" Other examples of queries within this genre are: "Is situation S improving or worsening?" and "Is agent A1 dominating or being dominated by agent A2?" This criterion can be thought of as *directionality*, in contrast to the *topicality* criterion of standard IR systems. Note that directional queries are domain independent.

How is this kind of classification to be accomplished? Clearly keyword-based analysis alone is not sufficient. Consider the classification criterion "Is the agent in favor of the event?" applied to the following pair of sentences:

(1a) Three congresswomen introduced legislation to *lift* the ban on the dumping of wastewater off of the coast.

(1b) Three congresswomen introduced legislation to *support* the ban on the dumping of wastewater off of the coast.

A difference of one word manages to reverse the attitude of the agents toward the situation, even though *lift* and *support* are not contradictory when out of context. To correctly distinguish these sentences, some sort of semantic component must come into play. However, the semantics need not be comprehensive – the interpretation mechanism can take advantage of the restricted nature of the query in order to minimize the degree of inference needed.

Task-based Text Interpretation

A mechanism that can classify an article based on the directionality criterion provides a precise interpretation of a narrow slice of the semantic content of the document. TTI, the method proposed to accomplish this, is founded on three observations:

1. IR can be used to isolate the documents that correspond to a topic, and keywords can be used to help pinpoint likely relevant sentences (Withgott *et al.* 1989).

2. A simple conceptual model can be used to describe some general semantic characteristics of a wide range of linguistic phenomena (Talmy 1985; Reddy 1979).

3. Ignoring semantic subtleties and interpreting lexical components in terms of general metaphoric descriptions can greatly simplify the mapping from syntax to the conceptual model.

TTI involves interpreting text in terms of a simple semantic model, only to the amount of detail necessary to accomplish the target task (to answer the query of interest). In effect, the process recasts a portion of the document into a predefined semantic model.

In brief, the method proceeds as follows: Relevant documents are selected by the system's Information Retrieval component, which makes use of domain dependent keywords and phrases (assumed to already be supplied) that identify the target concepts (e.g. the system knows about lexical items involved in expressing a topic such as "wastewater dumping"). This information is used to isolate sentences that are likely to contain the answer to the target query (e.g. a sentence that refers both to "congresswomen" and "wastewater dumping" is a good candidate). Once a candidate sentence is found, a partial parse is performed. As the analysis proceeds, pieces of the model are instantiated and linked together corresponding to elements of the parse. The resulting structure is interpreted in terms of the model, and the query is answered. This procedure is described in more detail below.

The Use of General Metaphor

Cognitive linguists such as Reddy (1979) and Lakoff & Johnson (1980) observe that the use of *general metaphors* is surprisingly widespread and consistent in "everyday" utterances. An example of a general English metaphor is one in which negative, undesirable things are described in terms of "downness." This is evident in phrases such as "take a dip," "the quality is declining," "it's going downhill," and so on. The central meaning of metaphors such as these can be considered to be domain independent, as evidenced by the fact that they are used in many diverse contexts.

A mechanism which makes inferences based solely on the central meanings of this kind of metaphor can exploit this property of domain independence. This is the main idea behind the semantic analysis component of TTI. If the task of interest can be described in terms that can be inferred from the central meaning of a metaphor, then the semantic analysis can be domain independent. As mentioned in the introduction, TTI uses this notion of metaphor in two distinct ways: both as a simple model into which the semantics of sentences

are coerced, and as a criterion to determine how to place words and phrases into lexical conglomerations, which are used by the coercion process.

The Semantic Model

The first idea, that of mapping sentence semantics into one basic metaphor, has been exemplified in the work of Talmy (1985) and Reddy (1979). Talmy, in his theory of force dynamics, uses an intuitive model to describe a multitude of linguistic phenomena in terms of a more general conceptual framework. Reddy describes how the *conduit* metaphor underlies many English expressions about communication. The main theme of this metaphor is that thoughts are objects which are placed by the speaker into containers that are sent along a conduit, and are removed from the containers at the other end by the listener. Inferences that can be made about conduits (e.g. they can be blocked up, become full, etc.) are applied to notions of communication as well (e.g. "Your meaning did not come through.").

Empirical analysis reveals that the inferences that can be generated based on a *path* metaphor suffice to answer the target query defined in an earlier section ("Is the agent in favor of the event?"). More specifically, this metaphor can be stated as "events are vehicles that travel along a path toward a destination." This can be thought of as the conduit metaphor augmented with a goal (the destination); the goal provides the necessary directional component. In this model, an entity is seen as progressing along a path from a starting point toward a destination. The entity may encounter barriers in its path, indicating that its tendency is being blocked. Agents independent of the entity have the power to introduce barriers, remove barriers, reinforce or weaken barriers, initiate the entity's journey, speed up or slow down the journey, or bring the entity to its destination.

In determining the answer to the target query, the directionality of two separate components must be determined – both the attitude of the agent and the progress of the event. This is necessary in order to determine whether or not an opinion has been expressed. In sentences such as "The congresswomen said that wastewater dumping is continuing," the system must realize that no agent is expressing an attitude. Furthermore, there is a class of expressions which indicates the direction of the agent's attitude relatively directly (e.g. "favors," "denounced"), and it is reasonable to take advantage of this. The attitude of the agent can be expressed indirectly, however, as in the phrase "introduced legislation," the interpretation of which is described in the next section.

Part of the power of this approach is that the set of valid heuristic inferences is limited and is explicitly specified with respect to the model. Some example heuristics for the model outlined above are:

- If an agent initiates an entity's journey along a path, the agent favors the entity's progress, and therefore is in favor of what the entity symbolizes.
- If one agent favors a measure that strengthens another agent's capability to remove a barrier from the path of an entity, the first agent favors the progress of the entity.

Preliminary work reveals that this semantic model, with some minor modifications, can be applied to answer another general query, namely "Does the event E improve the situation S?" This includes subquestions such as "Does the drug cure the disease?" and "Is the financial situation improving?" These queries all have a directional component. To be investigated still are queries that are best based on some other semantic model.

Both Talmy and Reddy consider the base metaphors that they investigate to be at least part of the underlying meaning of some subset of linguistic utterances. However, in TTI the base metaphor is used as a *lingua franca* into which the meanings of *all* candidate sentences are coerced. This is useful for two reasons: first, once the system has a representation of the sentence based on this metaphor, it need perform only a restricted set of inferences. Second, since the model being mapped into is small (compared with mapping into a network of "real-world" knowledge), the syntax-to-semantics conversion is simplified considerably.

Lexical Conglomerations

The second way in which TTI makes use of general metaphor is in the assignment of lexical items to conglomerations. Words such as "ban" and "roadblock" are easily assigned to the "barrier" conglomeration. However, assignment to conglomerations is not limited to grouping of what might usually be considered synonyms. For example, in the political arena, "limits" and "ceilings" are often proposed as compromises – alternatives to outright bans. If the system were to produce a detailed semantic interpretation, it would have to know how to reason about partially restricted movement. Although this is a valid approach, it is much simpler to coerce the notion of a limit into that of a barrier. This is justified by the heuristic that if A proposes a putting a limit on the amount of E, then A in reality is opposed to E but is suggesting a limit as a compromise. Thus "limit" would be placed in the conglomeration called "create-barrier."

As another example, consider the *war* metaphor. This may be manifested in sentences such as "The congresswoman attacked the bill on wastewater dumping," and "The president favors shielding the elephants against attacks by poachers." The verb "attack" is placed in the "create-barrier" conglomeration, since an attack on an entity can be viewed as an attempt to block its progress on a path. Similarly, "shielding" is placed in the "remove-barrier" conglomeration because shields are used to counteract the effects of attacks.[1]

All lexical items that can affect the direction of the outcome of the query must be classified. Once the basic conglomerations are defined, the relevant members of a lexicon must be annotated with the ID's of the conglomerations to which they belong. Many words are "opaque" with respect to the model – they do not require annotation and can be treated as "black boxes" without affecting the outcome of the analysis. Some others are "transparent" – they are used to link pieces of the model together but do not instantiate any part of the model.

How is conglomeration membership determined? Currently this is done by examining the contexts in which the lexeme occurs, considering the metaphors it appears in, its etymology, and how it interacts with the path model. It would be interesting to explore the possibility of automatic conglomeration assignment, as perhaps a simplified version of the word sense discovery task (Zernik 1989).

Differences from Other Metaphor-Based Approaches

Approaches such as (Carbonell 1982, Martin 1988) have incorporated the use of general, or conventional, metaphor in the context of general purpose text understanding. Not surprisingly, these approaches differ significantly from TTI's, since their goals are different (full interpretation versus classification along one dimension). Martin's system takes advantage of the structure underlying conventional metaphor to determine correspondences between source and target concepts, where "[t]he target concept is the concept that is actually under consideration. The source specifies the concepts in terms of which the target is being viewed" (pg. 30). For example, given the sentence "How do I kill a process?" the source domain is "killing," the target is "terminating," and the two are linked through an intermediate "terminate-as-killing" metaphor. This approach allows the system to make complex inferences that take into account detailed domain knowledge from both the source and the target. For example, the system must know enough about processes to know that it makes sense to discuss terminating one.

In TTI, by contrast, the notion of target and source are not used. Rather, every relevant lexical item is converted via the conglomerations into an element of the path domain. Inferences are performed only within the path domain (i.e., there is no equivalent to needing to represent the fact that processes are things that can be terminated). More fundamentally, TTI differs from these other approaches in that the latter attempt to interpret the metaphorical mappings in roughly the same way a reader would, whereas TTI exploits the regularities underlying the metaphors to coerce their meanings.

An Example

This section presents the exposition of a simple example, a modified version of sentence (1a):

(1c) Congresswomen introduced legislation to lift the ban on wastewater dumping.

The first step involves some simple syntactic analysis. For sentence (1c) this consists of: labelling of the part-of-speech categories, identification of target lexical units based on the keywords used in the original IR lookup (e.g. "congresswomen" and "wastewater dumping" are identified as target lexical units), and recognition of the predicates and what constitutes their complements. In the current system sentences are hand-parsed into a simple feature-structure representation.

The next step involves building an instance of the model from the parse. The main predicate, "introduced," is examined first; it indexes into an "initiate-event" conglomeration. Associated with each conglomeration is a template that represents the fragment of the model to which members of the conglomeration correspond. For example, associated with lexemes that can indicate the initiation of an event (e.g. "introduced," "began," "gave birth to") is a template: *initiate-event($Agent, $Event, $Purpose)*. The predicate's complements are examined, and since "congressman" is recognized as a target agent unit with the syntactic structure of a simple NP, it is instantiated into the *$Agent* slot.

At this point, an advantage of the TTI methodology becomes apparent: the system need not have a special interpretation for every lexical item it encounters. Due

[1] It may seem counterintuitive that the verb "shield" is used to indicate the *removal* of a barrier, since the noun shield often is thought of as a barrier of sorts. Actually, in this framework, the verb "shield" sometimes acts as the introduction of a barrier, and sometimes as the removal of one, depending, among other things, on the amount of information obtained from the prepositional phrase associated with it.

to the constraints imposed by the syntax of the sentence and the characteristics of this conglomeration, the analyzer need not have knowledge of how to interpret the simple NP "legislation," since no matter what its meaning, it cannot affect the directionality of the sentence. So, "legislation" is simply placed in the *$Event* slot. The remainder of the sentence builds up a fragment to fill in the (optional) *$Purpose* slot.

Based on its syntax and the kind of complements its predicate "introduce" is known to take, the infinitival "to lift" clause is interpreted as expressing the "purpose" of the action. The lexeme "lift" can index into a "remove-barrier" conglomeration, but in some cases this interpretation does not apply. The choice is delayed temporarily as the analyzer looks ahead to the next phrase, "the ban on wastewater dumping." The syntactic structure of two simple NP's linked by a preposition occurs often and is analyzed all at once. If the second NP is a target (as it is in this case), and the first NP is a nominalized form of a verb that indexes into a conglomeration (as is "ban"), then the system can use "ban's" conglomeration. Other phrases that have this structural pattern are "the creation of a new fund," "an all-out purge of selenium," and "other restrictions on gun ownership." In some cases the directionality of the first NP is partially or fully determined by the preposition that follows it. For example, "protest against" and "protest for" indicate opposing directions. As this example shows, some syntactic details must be preserved in the parse in order to ensure correct interpretation.

The analysis now has generated the fragment *initiate-barrier(A1, target([wastewater dumping]), P1)* (the agent slot remains uninstantiated; how this can be inferred is not discussed here). This fragment choice helps determine that the proper conglomeration for "lift" is indeed "remove-barrier." So the final template is: *initiate-event(target([congresswomen]), [legislation], remove-barrier(A0, initiate-barrier(A1, target([wastewater dumping]), P1), P0))*

This structure is now interpreted according to heuristics as described above, generating the inferences:

- the congresswomen favor legislation
- the congresswomen favor removing a ban
- the congresswomen favor wastewater dumping

Although the first two may not be salient, they are inferable from the model. The third inference answers the target query.

If sentence (1b) had been processed instead, "support" would have indexed into a conglomeration whose template indicates that its argument is "strengthened," rather than one indicating the removal of a barrier, as "lift" does. The interpretation heuristics would then encounter an instance of an agent (the congresswomen) initiating an event that *strengthens* a barrier-placing action, and would conclude that the congresswomen oppose the wastewater dumping.

The method is in the early stages of development: a prototype program that analyzes hand-chosen and hand-parsed sentences has been written in Common LISP. The most important next steps are showing that the method succeeds over a large set of documents and applying the method to a variety of direction-based queries.

Discussion

TTI is constrained by the goal of making semantic discernments while eschewing the complexity required by NLP systems that attempt to generate "all plausible" inferences. As mentioned in the introduction, this kind of approach is profitable only if the effort involved in building and executing the system does not outweigh the depth and quality of the results. If the effort does get too large, one could argue that a general text understanding system would be more appropriate (since it can produce more detailed interpretations), or that the semantic component should be scrapped altogether (because its results do not justify its cost).

In analyzing the "cost-benefit ratio" of an approach like TTI, several points of contention should be raised:

How valid is the assumption that the target queries are general and useful enough to justify the effort required to answer them? Note that this depends to some extent on how well the conglomeration information for one query applies to others.

How often does the text contain the answer to the query in a form discernible to the method? With newspaper articles this assumption is more plausible than it would be in some other discourse domains, but even so, the system will generate a partially complete classification. This is tolerable provided that the system does not make *incorrect* classifications and provided that the potential for errors of omission is known.

How much syntactic information must be incorporated into the conglomerations, i.e., to what extent do conglomerations have to be distinguished based on the syntactic context in which they appear? If this information becomes too detailed, the conglomerations may become numerous enough to counteract the motivation behind them.

The outcome of the tradeoffs can only be determined through empirical studies. If this approach and others like it can tip the balance in their favor, restricted

semantic analysis will occupy a dominant role in the construction of efficient, intelligent text interpretation systems.

Acknowledgements. The majority of this research was completed during an internship at Xerox Palo Alto Research Center. Thanks to Per-Kristian Halvorsen, John Batali, Susan Newman, Deborah Tater, Cathy Marshall, and Doug Cutting for much useful discussion in the preparation of this paper; additionally Robert Wilensky provided helpful comments on an earlier draft.

References

Carbonell, J. G. (1982). Metaphor: An inescapable phenomenon in natural-language comprehension. In W. G. Lehnert & M. H. Ringle, editors, *Strategies for Natural Language Processing*, chapter 15, pages 415–434. Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

Lakoff, G. & M. Johnson (1980). *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Martin, J. H. (1988). *A Computational Theory of Metaphor*. PhD thesis, University of California, Berkeley, Berkeley, CA.

McCume, B., R. Tong, J. Dean, & D. Shapiro (1985). Rubric: A system for rule-based information retrieval. *IEEE Transactions on Software Engineering*, 11(9).

Reddy, M. (1979). The conduit metaphor – a case of frame conflict in our language about language. In A. Ortony, editor, *Metaphor and Thought*, pages 284–324. University Press, Cambridge, England.

Talmy, L. (1985). Force dynamics in language and thought. In *Parasession on Causatives and Agentivity*, University of Chicago. Chicago Linguistic Society (21st Regional Meeting).

Withgott, M., F. Chen, D. Cutting, P.-K. Halvorsen, J. Kupiec, J. Pedersen, & J. Shrager (1989). Using emphasized information for indexing and retrieval. Technical report, Xerox Palo Alto Research Center, Internal Memo.

Zernik, U. (1989). Lexicon acquisition: Learning from corpus by capitalizing on lexical categories. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI. Morgan Kaufman Publishers, Inc.