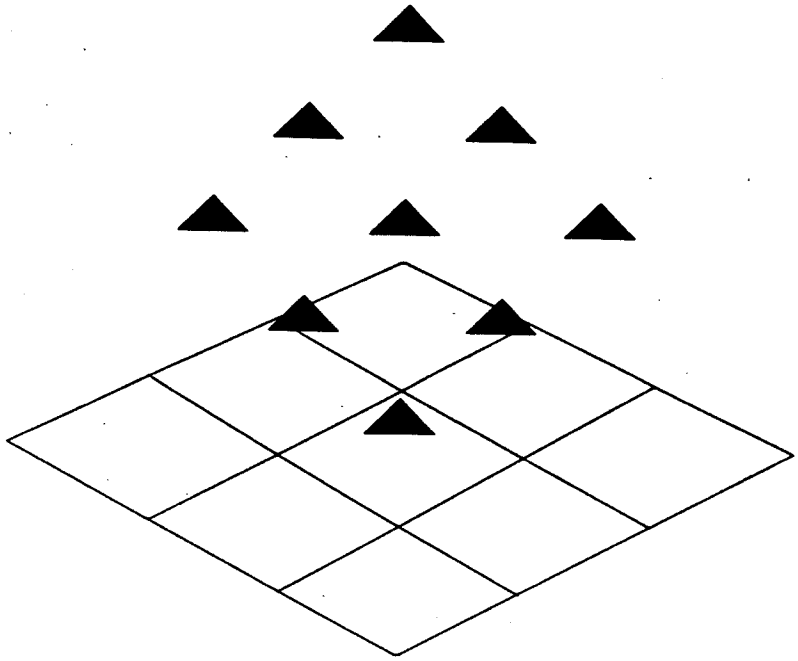
AAAI–94 Fall Symposium Series



Knowledge Representation for Natural Language Processing in Implemented Systems

Program Committee:
Syed S. Ali, Southwest Missouri State University, Chair
Douglas Appelt, SRI International
Lucja Iwanska, Wayne State University
Lenhart Schubert, University of Rochester
Stuart C. Shapiro, State University of New York at Buffalo

WORKING NOTES

November 4 – 6, 1994

The Monteleone Hotel, New Orleans, Louisiana

Relevance Reasoning in Text Retrieval

Kathleen Dahlgren
Intelligent Text Processing, Inc.
kd@itpinc.com

Introduction

Relevance in a text retrieval task concerns the relevance of texts to a natural language query. The most difficult aspect of the task is to render the natural language of the text and the query into a form which permits computational relevance reasoning. English (or any language) must be rendered into logic before logical reasoning can be applied to it. Once this has been accomplished, relevance reasoning becomes a problem of determining the similarity of object and event descriptions. The basic similarity assignment in computational relevance judgment is sensitive to several complexities: granularity, ellipsis, domain, and genre.

Textual Relevance

In the text retrieval task, relevance is a relation between two sets of propositions, those expressed by a query (Q) and those expressed by a library of texts (T). The assignment of relevance is gradient, in that propositions are more or less relevant to a query. A subset of T, T1, is relevant to Q if it contains descriptions of similar objects to those in Q, standing in similar, converse or antonymous relations as their relations in Q in appropropriate caseroles.

In psychology, similarity is a perceived relation between objects or events in the world. To mimic this task computationally would require the recognition of the similarity of events and objects by comparison of perceptual and functional features. It would require knowledge of the salience of certain features over others. However, in text retrieval, the problem is simplified because the authors of the query and text have predicated certain expressions of objects and events. Human predication categorizes objects as like other objects for which the same predicate would be used, so that objects are placed in classes of similar objects by predication. Objects described by natural language expressions have been assigned to similarity classes in advance. Thus, similarity in textual relevance is not analogy. The authors have judged which features of objects and events are most salient, and thereby which predications are most appropriate for them. The text retrieval task is to discover the similarity of predications, after the fact of human categorization and predication.

However people use a wide variety of expressions to describe similar events. Texts which should be judged relevant to a query may contain different predicates than the query does. Computational recognition of similarity requires prior knowledge of all the predications for similar objects and events. Predicates must be grouped into loose synonymy classes. Texts which are similar have predicates in the same synonymy classes describing objects which stand in similar, antonymous or converse relations, in appropriate caseroles.

Natural Language Interpretation

While textual relevance is simplified over the psychological problem of relevance, natural language presents difficult problems of interpretation. Superficially, natural language strings are both ambiguous and redundant. Words and structures which over are superficially identical have DIFFERENT meanings. On the other hand the SAME meaning can expressed redundantly in many different ways.

For example, DIFFERENT meanings look the same in the following sentences:

John took the boat up the river.

John took up boating on the river.

John took down the mast on the boat by the river.

For a pattern-matching algorithm, all of these sentences contain the substance words of a query "Did John take the boat up the river?". The English word "take" has forty-eight different senses, and each of these sentences has a different sense of "take". Only one sense is relevant to the query. In a text retrieval system which fully analyzes each sentence syntactically and semantically, words can be disambiguated. The disambiguated words are given different predicates in the cognitive model of the text, so that a text about

physically taking will only be relevant to those cognitive models with the physical reading of "take" or similar predicates. .

On the other hand, the SAME meaning has the potential of being expressed in hundreds of different ways, as in the following:

1) John steered the boat up the river.

2) John's boat trip was up the river.

3) John cruised up the river.

4) The boat travelled up the river. John was captain.

5) John was captain of a boat. He took it up the river.

The first two text is a basic active sentence in which there is a verb (steer), subject (John) and object (the boat) and prepositional phrase (up the river). It is relatively straightforward to assign relevance for text (1) and a query "Did John take the boat up the river?" by recognizing the similarity of the predicates "take" and "steer". The second text expresses the same event with a nominal "trip". To find that text (2) is relevant to the query "Did John take the boat up the river?", the relevance mechanism must find the subject and object cases of the nominal, and then reason that subject and object case roles are similar to subject and object cases of the query verb. The verb in third text has a single word "cruise" to describe the event of "taking a boat" in the query. The relevance mechanism must see the similarity of single with multiple predicates. The fourth text has identifies John with an agentive role "captain". To recognize that John is captain in the "travel" event of the prior sentence, the two sentences have to be related by coherence reasoning. The fifth text has pronouns which must be resolved in order to find that John took the boat.

Intelligent Text Processing's full natural language approach to text retrieval has modules which imitate every level of human linguistic reasoning. The propositions of the text and query are produced when a natural language understanding system, Interpretext, translates English into logic. It parses the text [E. P. Stabler, 1992], disambiguates words and structures [Dahlgren, McDowell, & Stabler, 1989], forms a discourse representation structure [Wada & Asher, 1986], finds antecedents of anaphoric expressions [Wada, 1994], and assigns coherence relations between events [Dahlgren, 1989]. The translation of "John took the boat up the river" is, in part, as follows.

```
take3(e1,n1,the1).
boat1(t1,the1).
anchor(n1,john).
direction(e1,the2).
river(the2).
before(e1,now).
```

Interpretext retrieves and uses world knowledge in the form of naive semantic representations for the content words in the text [Dahlgren, 1988]. The lexical representations are naive theories of the properties of objects and the implications of events, and are available during relevance reasoning. Megabytes of text in seven domains have been interpreted by this system. Text retrieval in one domain has been implemented and exceeds 95% precision and recall using the methods described in this paper [Dahlgren, 1992].

Since Interpretext translates text into a cognitive model consisting of logic expressions, with all words disambiguated and all anaphoric expressions resolved, extremely precise textual relevance reasoning is possible. In contrast with pattern-matching techniques, this form of representation prevents false hits for cases where the target text has many of the same strings as the query, used in different senses. For example, for our query "Did John take the boat up the river?", Interpretext with full disambiguation does NOT retrieve the text "John took up boating on the river", because a different meaning of "take" is used in the query and target text. This full form of representation is not fooled when argument structures differ. The query, "Did John take Mary?" it will not retrieve texts about Mary taking John. Because anaphors are resolved, it can find a text with a pronoun or other anaphoric expression in place of a query word, as in text (5) above.

Relevance of Objects

Relevance judgement differs for objects, events and situations. Object similarity in textual relevance can be determined with simple predicate matching and loose synonymy, or by more complex reasoning with background knowledge. Simple predicate matching relies upon word sense disambiguation. In the following texts, "train" has two different senses. Only T1, where "train" has a vehicle sense, is relevant to the query Q.

T1. John took a train up the river.

T2. John traveled with his followers in train.

Q. Did John travel by train?

The cognitive model of T1 is, in part,

```
take3(e1,n1,a1).
train1(a1).
direction(e1,the1).
river(the1).
```

The cognitive model of Q is, in part,

travel1(e2,n2).
means(e2,a2).
train1(a2).

In the similarity computation, take3 and travel1 are found to be similar event predications, train1 and train1 found to be similar object predications, and the caserole "means" is found to fulfill the same role as the object role of take3.

However, the cognitive model of T2 is, in part,

travel1(e2,n2).
accompaniment(e2,c1).
follower(c1).
location(c1,the3).
train5(the3).

In the similarity computation, travel1 and travel1 are found to be similar events predications, but train1 and train5 are found to be dissimilar object predications, and relevance fails.

Object names or their synonyms may not be present in a relevant text. An object with similar functions can be relevant ("hoe" is relevant to "trowel", "scalpel" to "knife"). An instrumental object can stand for an activity or role ("hoe" for "gardening", "scalpel" for "surgery" or "surgeon". Metonymic relations can drive relevance assignment ("mainsail" for "boat", "tractor" for "farm" or "farming"). A textual object which falls into the class of objects named by a query predicate is relevant (Q has "boat" and T has "trawler"; Q has "river" and T has "Amazon"). The naive semantic lexical representations in Interpretext contain these synonyms, functions, instruments, part and ontological relations, so that relevance reasoning requires no new types of lexical information, though relevance reasoning drives lexical enhancement.

Beyond simple predicate similarity, more complex relations must be found. An expression with several predicates has the same meaning as (denotes the same kinds of objects as) a single predicate, so that paraphrase relations must be recognized (e.g., "fishing boat" and "trawler"). On the other hand, parts of expressions are redundant or empty and must be ignored ("factor" in "risk factor", "have" in "boat which has a jib").

Relevance of Events

Relevant textual events are those which involve similarly described objects in similarly described relations in appropriate caseroles. Similarity of events, as with objects, can be assigned through loose synonymy of predicates ("travel" and "journey") or through reasoning with world knowledge. Given a full text interpretation with argument structure, the relevance reasoner rejects texts with the same objects in converse relations . It knows the difference between "John took Mary on the trip" and "Mary took John on the trip". It also knows that "John took Mary on the trip" is similar to "Mary was taken on the trip by John", and "Mary went on John's trip". Events with similar consequences are similar ("travel", "drive", "come"). Converse events are similar as long as arguments are in the appropriate caseroles ("leave" and "return", "receive" and "give"). Sometimes the function of an object makes it relevant to an event ("book" to "read", "tractor" to "ploughing", "knife" to "cutting"). Antonyms are relevant ("win" and "lose"). If an event is queried, and its negation has been reported in a text, the text is typically relevant.

Both verbs and nominals are predicated of events. T with "trip up the river" is relevant to Q with "travel up the river". The recovery of caseroles in nominals is important for accurate retrieval. For example, text (6) is relevant to query (8), but not text (7).

(6) John's trip up the river was fun.

(7) The trip up the river for John was fun.

(8) Who went up the river?

This can be discovered by inspection of caseroles of the nominal "trip". The cognitive model of (6), has the following predicates, in part, showing John as subject of the trip.

trip1(e1).
subject(e1,n1).
anchor(n1,john).
direction(e1,the2).
river(the2).
fun(e1).

In contrast, the cognitive model of (7) has a benefactive caserole for John, and no subject for the trip. So it is irrelevant query (8).

trip1(e2).
benefactive(e2,n1).
anchor(n1,john).
direction(e2,the2).
river(the2).
fun(e2).

Relevance of Situations

The relevance reasoning mechanisms presented in this paper are inadequate for queries which describe complex situations with many predicates. They are sufficient to produce retrievals as long as queries are relatively short (under ten verbs). For longer queries, a mechanism to compute the salience of events in the query should be added. Otherwise, no relevant texts can be found, because the probability of finding any text describing all of the objects in all of the relations of the situation described by the long query approaches zero.

Complexities of Textual Relevance Reasoning

One of the complexities in relevance reasoning is ontological granularity of the query, individual texts, and the library. Queries which are too broad must be recognized and intercepted. A query about "Latin America" to a library sufficiently large to have an unmanageable number of retrievals about Latin America is too broad. A query needs to be at about the granularity of some text topic, segment topic or sentence, and not broader. As overly narrow queries fail, no screening against them is needed. In general, if Q is coarser than T, and the predicates of T are below Q in the ontology, it is relevant. If Q is finer-grained than T, and propositions similar to Q are in T, T is relevant.

Another complexity is anaphora and ellipsis. The natural language interpreter must recover antecedents of linguistic devices such as pronouns and definite descriptions. Otherwise recall is lowered by failure to assign relevance to elements which stand for relevant predicates. In (5), "he" stands for John, and "it" stand for the boat, so that (5) is relevant to the Q "Did John take the boat up the river?". In order to compute the relevance of (5), these pronouns must be resolved.

Furthermore, assuming the implicit relevance of topics predicates can enhance recall. In a text entitled "John's cruise up the river", John, the boat and the river are implicitly present and relevant throughout the text. Similarly, events in a segment of text are potentially relevant to all events in the same segment, but not to events in other segments.

A third complexity is domain, or field of activity. Texts in a certain domains prefer particular word senses. For example, in financial texts the financial senses of "bank" and "asset" are preferred, while in geography texts, other senses are preferred.

A final complexity is genre. In highly structured texts, especially those with explicitly formulated requirements, the automatic extraction of topics and segments is not necessary because the authors are required to segment the text and provide titles of sections. Using segments, the search for relevant events can be speeded up. Reasoning to less structured texts is slower because it cannot be focused in segments. Location of antecedents of anaphors is less precise in less structured texts, as well, resulting in false hits.

Conclusion

In conclusion, precise textual relevance reasoning requires disambiguated words, recovery of argument structure, assignment of case roles, interpretation of nominals and reference to world knowledge. Relevance reasoning is enhanced by treatment of ellipsis, and recognition of the domain of a text. To the extent that the query and text have similar granularity, or the query is narrower, the relevance reasoning can be more successful. To the text extent that the text is explicitly structured, the text interpreter can more accurately recover antecedents and topics, and the relevance reasoning can thereby be more precise and have better recall.

References

Dahlgren, K.; McDowell, J. P.; and Stabler, E. P., J. 1989. Knowledge representation for commonsense reasoning with text. *Computational Linguistics.*

Dahlgren, K. 1988. *Naive Semantics for Natural Language Understanding.* Boston, MA: Kluwer.

Dahlgren, K. 1989. Coherence relation assignment. In *Proc. Cognitive Science Society.*

Dahlgren, K. 1992. Interpretation of textual queries using a cognitive model. In Lauer, T.; Peacock, E.; and Graesser, A., eds., *Questions and Information Systems.* Hillsdale, NJ: Erlbaum.

E. P. Stabler, J. 1992. *The Logical Approach to Syntax: Foundations, Specifications, and Implementations of Theories of Government and Binding.* Cambridge, MA: MIT Press.

Wada, H., and Asher, N. 1986. BUILDRS: An implementation of DR theory and LFG. In *Proc. COLING,* 540–545.

Wada, H. 1994. A treatment of functional definite descriptions. In *Proc. COLING.*