Using Coreference to Improve Passage Retrieval for Question Answering

Thomas S. Morton
Department of Computer and Information Science
University of Pennsylvania
tsmorton@linc.cis.upenn.edu

Abstract

We present a system which retrieves answers to queries based on term weighting supplemented by coreference relationships between entities in a document. An evaluation of this system is given which demonstrates that the coreference relationships allow significantly more questions to be answered than with a baseline system which doesn't model these relationships.

Introduction[1]

Search engines have become ubiquitous as a means for accessing information. When a ranking of documents is returned by a search engine the information retrieval task is usually not complete. The document, as a unit of information, is often too large for many users information needs and finding information within the set of returned documents poses a burden of its own. Here we examine a technique for extracting sentences from documents which attempts to satisfy the users information needs by providing an answer to the query presented. The system does this by modeling coreference relationships between entities in the document and uses this information to better select passages which contain answers. An evaluation of this system is given which demonstrates that it performs better than a standard weighting scheme without coreference relations.

Problem Statement

A query indicates an informational need by the user to the search engine. The information required may take the form of a sentence or even a noun phrase. Here the task is to retrieve the passage of text which contains the answer to the query from a small collection of documents. Passages are then ranked and presented to the user. We only examine queries to which answers are likely to be stated in a sentence or noun phrase since answers which are typically longer can be difficult to annotate reliably. This technology differs from the standard document ranking task in that, if successful the user will likely not need to examine any of the retrieved documents in their entirety. This also differs from the document summarization. provided by many search engines today, in that the sentences selected are influenced by the query and are selected across multiple documents.

We view a system such as ours as providing a secondary level of processing after a small set of documents, which the user believes contain the information desired, have been found. This first step would likely be provided by a traditional search engine, thus this technology serves as an enhancement to an existing document retrieval systems rather than a replacement. Advancements in document retrieval would only help the performance of a system such as ours as these improvements would increase the likelihood that the answer to the user's query is in one of the top ranked documents returned.

Approach

One of the difficulties in extracting text from the middle of a document is that this text refers to its surrounding discourse context and often is incomplete without that context. While the relationship between a sentence and the discourse context in which it appears can be extremely complex, some of these relationships can be modeled. In this paper we attempt to model identity coreference relationships between noun phrases in a document and then use these relationships to better rank segments of text for question answering.

The system currently models coreference relationships between entities occuring as proper noun phrases, such as names and locations, definite noun phrases, typically beginning with "the" in English, and non-possessive, third person pronouns such as "he", "she", "it", and "they".

Implementation

The system processes a query and a set of documents by first preprocessing the documents to prepare them for the coreference component of the system. Once this is done the system identifies coreference relationships between noun phrases in the document. Finally each

[1] Copyright ©1999, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

sentence is ranked based on its terms and the terms of noun phrases coreferent with it.

Preprocessing

Determining coreference between noun phrases requires that the noun phrases in the text have been identified. This processing begins by preprocessing the HTML to determine likely boundaries between segments of text, sentence detecting these segments using a sentence detector described in (Reynar & Ratnaparkhi 1997), and tokenizing those sentences using a tokenizer described in (Reynar 1998). The text can then be part of speech tagged using the tagger described in (Ratnaparkhi 1996) and finally noun phrases are determined using a maximum entropy model trained on the Penn Treebank(Marcus, Santorini, & Marcinkiewicz 1994). The output of Nymble(Bikel *et al.* 1997), a named entity recognizer which determines which words are people's names, organizations, locations, etc., is also used for determining coreference relationships.

Coreference

Once preprocessing is completed the system iterates through each of the noun phrases to determine if they refer to a noun phrase which has occured previously. Only proper noun phrases, definite noun phrases, and non-possessive third person pronouns are considered. Proper nouns phrases are determined by the part of speech assigned to the last word in the noun phrase. A proper noun phrase is considered coreferent with a previously occuring noun phrase if it is a substring of that noun phrase excluding abbreviations and words which are not proper nouns. A noun phrase is considered definite if it begins with a the determiner "the" or begins with a possessive pronoun or a past participle verb. A definite noun phrase is considered coreferent with another noun phrase if the last word in the noun phrase matches the last word in a previously occuring noun phrase. The mechanism for resolving pronouns consists of a maximum entropy model which examines two noun phrases and produces a probability that they co-refer. The twenty previously occuring noun phrases are considered as well as the possibility that the pronoun refers to none of these noun phrases. The pair with the highest probability are considered coreferent or the pronoun is left unresolved when the model predicts that the most likely outcome is that the pronoun doesn't refer to any of the proceeding noun phrases. The model considers the following features:

1. The category of the noun phrase being considered as determined by the named entity recognizer.
2. How many noun phrases occur between the candidate noun phrase and the pronoun.
3. How many sentences occur between the candidate noun phrase and the pronoun.
4. Which noun phrase in a sentence is being referred to (first, second, ...).
5. In which noun phrase in a sentence did the pronoun occur (first, second,...).
6. Which pronoun is being considered.
7. Are the pronoun and the noun phrase compatible in number.
8. If the candidate noun phrase is another pronoun is it compatible with the referring pronoun.
9. If the candidate noun phrase is another pronoun is it the same as the referring pronoun.

The model is trained on nearly 1200 annotated examples of pronouns which refer or fail to refer to previously occurring noun phrases.

Segment Ranking

Segments for both systems are scored and ranked based on the sum of the *idf* weights(Salton 1989) for each unique term which occurs in the segment and also occurs in the query. The *idf* weights are computed based on the documents found on TREC discs 4 and 5 (Voorhees & Harman 1997). No additional score is given for tokens occuring more than once in a segment.

Segments for the baseline system are 250 byte sequences of text. The coreference enhanced system ranks each sentences detected during the preprocessing phase. In the coreference enhanced system, terms which occur in a coreferential noun phrase are also considered to have occured in that segment.

Evaluation

For the evaluation of the system fifty queries were selected from a collection of actual queries presented to an online search engine[2]. Queries were selected based on their expressing the users information need clearly, their being likely answered in a single sentence, and non-dubious intent.

The queries were then presented to Altavista[3] and the top 20 ranked documents were retrieved. The documents were then processed by the baseline system as well as the coreference system and the top ten ranked text segments were presented to the user. Since some of the documents retrieved were very long the baseline system was used to select the best segment and then the coreference system was applied to a 10K window surrounding that segment.

Each system presented the user with the top 10 ranked segments. Segments for both systems were limited to 250 bytes. In the coreference system equal amounts of text from each side of the sentence were added if the sentence was less than 250 bytes and the sentence was truncated if it was more than 250 bytes. The output of the systems was evaluated by a novice user and the systems were scored using the following metric:

[2] Electric Monk, `www.electricmonk.com`

[3] Altavista, `www.altavista.com`

$$\frac{\sum_{q=1}^{n} \frac{10-(rank(q)-1)}{10}}{n}$$

Here n is the number of questions processed and $rank(q)$ is the ranking of the first segment to answer question q. Using this metric the systems performed as follows:

Baseline	Coreference
30.66%	44.89%

Discussion

Text extraction and ranking while similar in its information retrieval goals with document ranking appears have very different properties. While a document can often stand alone in its interpretation the interpretation of a sentence is dependent on the context in which it appears. The modeling of the discourse context gives the coreference based system an advantage over a baseline model in situations where referring expressions, which provide little information outside of their discourse context, can be related to the query. The most extreme example of this being the use of pronouns.

In some cases the baseline system had a segment which answered the question ranked higher than the coreference system. This is because a 250 byte segment might span multiple sentences where the coreference system ranked single sentences. However, for all questions which were answered by the baseline system the coreference system also provided an answer. Overall the baseline was able to provide answers to 17 of the 50 questions and the coreference system answered 26. Since no attempt was made to ensure that an answer existed in the retrieved document collection it is likely that for some questions none of the top 20 documents retrieved contained an answer.

Related Work

Other work employs coreference for document summarization such as (Baldwin & Morton 1998) for single documents and (Mani & Bloedorn 1997) for multiple documents. These works differ in that they evaluate summaries based on a users ability to determine a documents relevance to a query rather than a specific information need. (Hirschman *et al.* 1999) reports improved performance on a system to take standardized reading comprehension tests for 3rd to 6th graders when coreference relationship were included.

Conclusion

Our system demonstrates that the automatic recognition of coreference relationships in text can improve performance in passage retrieval for question answering. The system has been tested on text retrieved from the World Wide Web and performed substantially better than the a baseline system which did not model coreference. In the future we would like to extend the scope of the coreference relations which are recovered from the text. Specifically we would like to model quoted speech contexts which would allow us to consider first and second person pronouns.

References

Baldwin, B., and Morton, T. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing.*

Bikel, D.; Miller, S.; Schwartz, R.; and Weischedel, R. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing.*

Hirschman, L.; Light, M.; Breck, E.; and Burger, J. D. 1999. Deep read: A reading comprehension system. In *Proceeding of the 37th Annual Meetin of the Association for Computational Linguistics.*

Mani, I., and Bloedorn, E. 1997. Multi-document summarization by graph search and matching. In *Proceeding of the Fourteenth National Conference on Artificial intelligence (AAAI-97).*

Marcus, M.; Santorini, B.; and Marcinkiewicz, M. 1994. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics* 19(2):313–330.

Ratnaparkhi, A. 1996. A Maximum Entropy Part of Speech Tagger. In Brill, E., and Church, K., eds., *Conference on Empirical Methods in Natural Language Processing.*

Reynar, J. C., and Ratnaparkhi, A. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 16–19.

Reynar, J. 1998. *Topic Segmentation: Algorithms and Applications.* Ph.D. Dissertation, University of Pennsylvania.

Salton, G. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Addison-Wesley Publishing Company, Inc.

Voorhees, E. M., and Harman, D. 1997. Overview of the fifth Text REtrieval Conference (TREC-5). In *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, 1–28. NIST 500-238.