Using Coreference for Question Answering

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

Abstract

We present a system which retrieves answers to queries based on coreference relationships between entities and events in the query and documents. An evaluation of this system is given which demonstrates that the the amount of information that the user must process on average, to find an answer to their query, is reduced by an order of magnitude.

1 Introduction

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 and events in the query and documents. An evaluation of this system is given which demonstrates that it performs better than using a standard $tf \cdot idf$ weighting and that the amount of information that the user must process on average, to find an answer to their query, is reduced by an order of magnitude over document ranking alone.

2 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. Sentences 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 are 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.

3 Approach

A query is viewed as identifying a relation to which a user desires a solution. This relation will most likely involve events and entities, and an answer to this relation will involve the same events and entities. Our approach attempts to find coreference relationships between the entities and events evoked by the query and those evoked in the document. Based on these relationships, sentences are ranked, and the highest ranked sentences are displayed to the user.

The coreference relationships that are modeled by this system include identity, part-whole,

and synonymy relations. Consider the following query and answer pairs.

> *Query:* What did Mark McGwire say about child abuse?
>
> *Sentence:* "What kills me is that you know there are kids over there who are being abused or neglected, you just don't know which ones" McGwire says.

In the above query answer pair the system attempts to capture the identity relationship between *Mark McGwire* and *McGwire* by determining that the term *McGwire* in this sentence is coreferent with a mention of *Mark McGwire* earlier in the document.. This allows the system to rank this sentence equivalently to a sentence mentioning the full name. The system also treats the term *child abuse* as a nominalization which allows it to speculate that the term *abused* in the sentence is a related event. Finally the verb *neglect* occurs frequently within documents which contain the verb *abuse*, which is nominalized in the query, so this term is treated as a related event. The system does not currently have a mechanism which tries to capture the relationship between *kids* and *children.*

> *Query:* Why did the U.S. bomb Sudan?
>
> *Sentence:* Last month, the United States launched a cruise missile attack against the Shifa Pharmaceutical Industries plant in Khartoum, alleging that U.S. intelligence agencies have turned up evidence – including soil samples – showing that the plant was producing chemicals which could be used to make VX, a deadly nerve gas.

In this example one of the entity-based relationships of interest is the identity relationship between *U.S.* and *United States.* Also of interest is the part-whole relationship between *Sudan* and *Khartoum*, it's capital. Finally the *bomb* event is related to the *launch/attack* event. The system does not currently have a mechanism which tries to capture the relationship between *Why* and *alleging* or *evidence.*

4 Implementation

The relationships above are captured by a number of different techniques which can be placed in essentially two categories. The first group finds identity relationships between different invocations of the same entity in a document. The second identifies more loosely defined relationships such as part-whole and synonymy. Each of the relationships identified is given a weight and based on the weights and relationships themselves sentences are ranked and presented to the user.

4.1 Identity Relationships

Identity relationships are first determined between the string instantiations of entities in single documents. This is done so that the discourse context in which these strings appear can be taken into account. The motivation for this comes in part from example texts where the same last name will be used to refer to different individuals in the same family. This is often unambiguous because full names are used in previous sentences, however this requires some modeling of which entities are most salient in the discourse. These relations are determined using techniques described in (Baldwin et al., 1998).

Another source of identity relationships is morphological and word order variations. Within noun phrases in the query the system constructs other possible word combinations which contain the head word of the noun phrase. For example a noun phrase such as "the photographed little trouper" would be extended to include "the photographed trouper", "the little trouper", and "the trouper" as well as variations excluding the determiner. Each of the variations is given a weight based on the ratio of the score that the new shorter term would have received if it had appeared in the query and the actual noun phrase that occured. The morphological roots of single word variations are also added to the list a possible terms which refer to the entity or event with no additional deduction in weighting. Finally query entities which are found in an acronym database are added to the list of corefering terms as well with a weight of 1.

4.2 Part-Whole and Synonymy Relationships

The system captures part-whole and synonymy relationships by examining co-occurrence statistics between certain classes of words. Specifically co-occurrence statistics are gathered on verbs and nominalization which co-occur much more often then one would expect based on chance alone. This is also done for proper nouns. For each verbal pair or proper noun pair the mutual information between the two is computed as follows:

$$I(w_1, w2) = log(\frac{p(w_1, w_2)}{p(w_1)p(w_2)})$$

where w_1 and w_2 are words and an event is defined as a word occuring in a document. All words $w2$ for which $I(w_1, w_2)$ exceeds a threshold where w_1 is a query term are added to the list of terms with which the query term can be referred to. This relationship is given with a weight of $I(w_1, w_2)/N$ where N is a normalization constant. The counts for the mutual information statistics were gathered from a corpus of over 62,000 Wall Street Journal articles which have been automatically tagged and parsed.

4.3 Sentence Ranking

Before sentence ranking begins each entity or event in the query is assigned a weight. This weight is the sum of inverse document frequency measure of the entity or events term based on its occurrence in the Wall Street Journal corpus described in the previous section. This measure is computed as:

$$idf(w_1) = log(\frac{N}{df(w_1)})$$

where N is the total number of documents in the corpus and $df(w_1)$ is the number of documents which contain word w_1. Once weighted, the system compares the entities and events evoked by the query with the entities and events evoked by the document. The comparison is done via simple string matching against all the terms with which the system has determined an entity or event can be referred to. Since these term expansions are weighted the score for for a particular term w_2 and a query term w_1 is:

$$s(w_1, w_2) = idf(w_1) \times weight_{w1}(w_2)$$

where $weight_{w1}$ is the weight assigned during one of the previous term expansion phases and idf is defined above. The $weight_{w1}$ function is defined to be 0 for any term w_2 for which no expansion took place. The score for the a particular entity or event in the document with respect to an entity or event in the query is the maximum value of $s(w_1, w_2)$ over all values of w_1 and w_2 for that entity or event. A particular sentence's score is computed as the sum of the scores of the set of entities and events it evokes.

For the purpose of evaluation a baseline system was also constructed. This system followed a more standard information retrieval approach to text ranking described in (Salton, 1989). Each token in the the query is assigned an idf score also based on the same corpus of Wall Street Journal articles as used with the other system. Query expansion simply consisted of stemming the tokens using a version of the Porter stemmer and sentences were scored as a sum of all matching terms, giving the familiar $tf \cdot idf$ measure.

5 Evaluation

For the evaluation of the system ten queries were selected from a collection of actual queries presented to an online search engine. 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 used in this evaluation are as follows:

- Why has the dollar weakened against the yen?
- What was the first manned Apollo mission to circle the moon?
- What virus was spread in the U.S. in 1968?
- Where were the 1968 Summer Olympics held?
- Who wrote "The Once and Future King"?
- What did Mark McGwire say about child abuse?
- What are the symptoms of Chronic Fatigue Syndrome?
- What kind of tanks does Israel have?
- What is the life span of a white tailed deer?

- Who was the first president of Turkey?

The information requested by the query was then searched for from a data source which was considered likely to contain the answer. Sources for these experiments include Britannica Online, CNN, and the Web at large. Once a promising set of documents were retrieved, the top ten were annotated for instances of the answer to the query. The system was then asked to process the ten documents and present a ranked listing of sentences.

System performance is presented below as the top ranked sentence which contained an answer to the question. A question mark is used to indicate that an answer did not appear in the top ten ranked sentences.

Query	First answer's rank	
	Full System	Baseline
1	2	4
2	2	3
3	8	6
4	2	4
5	7	8
6	1	3
7	4	?
8	?	?
9	1	1
10	1	1

6 Discussion

Sentence 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 very dependent on the context in which it appears. The modeling of the discourse gives the entity based system an advantage over a token based models in situations where referring expressions which provide little information outside of their discourse context can be related to the query. The most extreme example case of this being the use of pronouns.

The query expansion techniques presented here are simplistic compared to many used in for information retrieval however they are trying to capture different phenomenon. Here the goal is to capture different lexicalizations of the same entities and events. Since short news articles are likely to focus on a small number of entities and perhaps a single event or a group of related events it is hoped that the co-occurrence statistics gathered will reveal good candidates for alternate ways in which the query entities and events can be lexicalized.

This work employs many of the techniques used by (Baldwin and Morton, 1998) for performing query based summarization. Here however the retrieved information attempts to meet the users information needs rather then helping the user determine whether the entire document being summarized possibly meets that need. This system also differs in that it can present the user with information from multiple documents. While query sensitive multi-document systems exist (Mani and Bloedorn, 1998), evaluating such systems for the purpose of comparison is difficult.

Our evaluation shows that the system performs better than the baseline although the baseline performs surprisingly well. We believe that this is, in part, due to the lack of any notion of recall in the evaluation. While all queries were answered by multiple sentences, for some queries such as 4,5 and 10 it is not clear what benefit the retrieval of additional sentences would have. The baseline benefited from the fact that at least one of the answers typically contained most of the query terms. Classifying queries as single answer or multiple answer, and evaluating them separately may provide a sharper distinction in performance.

Comparing the users task with and without the system reveals a stark contrast in the amount of information needed to be processed. On average the system required 290 bytes of text to display the answer to the query to the user. In contrast, had the user reviewed the documents in the order presented by the search engine, the answer on average, would appear after more than 3000 bytes of text had been displayed.

7 Future Work

As a preliminary investigation into this task many areas of future work were discovered.

7.1 Term Modeling

The treatment of entities and events needs to be extended to model the nouns which indicate events more robustly and to exclude relational

verbs from consideration as events. A probabilistic model of pronouns where referents are treated as the basis for term expansion should also be considered. Another area which requires attention is wh-words. Even a simple model would likely reduce the space of entities considered relevant in a sentence.

7.2 Tools

In order to be more effective the models used for basic linguistic annotation, specifically the part of speech tagger, would need trained on a wider class of questions than is available in the Penn Treebank. The incorporation of a Name Entity Recognizer would provide additional categories on which co-occurrence statistics could be based and would likely prove helpful in the modeling of wh-words.

7.3 User Interaction

Finally since many of the system's components are derived from unsupervised corpus analysis, the system's language models could be updated as the user searches. This may better characterize the distribution of words in the areas the user is interested which could improve performance for that user.

8 Conclusion

We have presented a system which ranks sentences such that the answer to a users query will be presented on average in under 300 bytes. This system does this by finding entities and events shared by the query and the documents and by modeling coreference relationships between them. While this is a preliminary investigation and many areas of interest have yet to be explored, the reduction in the amount of text the user must process, to obtain the answers they want, is already dramatic.

References

Breck Baldwin and Thomas Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing*, Granada, Spain, June.

B. Baldwin, T. Morton, Amit Bagga, J. Baldridge, R. Chandraseker, A. Dimitriadis, K. Snyder, and M. Wolska. 1998. Description of the UPENN CAMP system as used for coreference. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Baltimore, Maryland.

Inderjeet Mani and Eric Bloedorn. 1998. Machine learning of generic and user-focused summarization. In *Proceeding of the Fifteenth National Conference on Artificial intelligence (AAAI-98)*.

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