Seeking an Upper Bound to Sentence Level Retrieval in Question Answering

Kieran White
Dept. of Computer Science
and Information Systems
University of Limerick
Limerick Ireland
Kieran.White@ul.ie

Richard F. E. Sutcliffe
Dept. of Computer Science
and Information Systems
University of Limerick
Limerick Ireland
Richard.Sutcliffe@ul.ie

ABSTRACT

This article documents a study investigating the specific requirements of a Question Answering System which relies on sentence level document retrieval. We compared 50 TReC factoid queries with the Aquaint document sentences which contain their answers. The main morphological and semantic relations linking query terms to document terms are identified. Hypernyms and simple co-occurrence mappings between query and document terms are found to be surprisingly important. We also determine that reasonably high accuracy can be attained by a Question Answering System which assesses each document sentence individually.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*query formulation, selection process*

General Terms

Theory, Verification

1. INTRODUCTION

The introduction of Question Answering (QA) to TReC in 1999 [7] was responsible for the current spate of activity in the field. Most QA systems which have been developed work in a similar fashion: First the query's type is identified, then an appropriate search expression is presented to a retrieval system which returns a set of documents. The documents are processed and a named entity of the desired type is returned.

The authors' previous work [6] involved the DLT QA System which operates in the same manner. In TReC 2003 DLT returned a set of paragraphs in response to a boolean search request. This paper concerns a refinement to this approach using sentence level indexing and retrieval.

Sentence level retrieval has much to recommend it, not least that one can be assured that search terms in the returned sentences will be in close proximity to one another, increasing the likelihood that any named entity eventually returned will be closely related to the subject of the query.

However this technique gains simplicity in the latter stages of the QA process in exchange for the requirement that the best possible use be made of every query term during retrieval. The QA system cannot rely on document verbosity to compensate for incompatibly worded queries. If a query term is in some way related to a sentence in the collection that fact must be recognised during retrieval. There is also the danger with sentence level retrieval that insufficient information is available in a single sentence to determine whether it answers the query.

The objective of this paper is to determine the types of morphological and semantic transformations necessary to associate query terms with related sentence terms by manually comparing a subset of TReC queries with answer bearing document sentences from the Aquaint collection. By recognising related terms we hope to increase recall during the retrieval stage. Some of the simpler morphological relationships can be handled by stemming. Monz [2] has already concluded that there are substantial gains in performance associated with stemming in the context of QA. Other types of relations can be identified in a running QA system with Predictive Annotation [4] which can tag document terms with morphological or semantic information during indexing. We envisage taking this approach to implement the ideas discussed later. A semantic analysis of terms could also be used to increase precision. Pasca [3] refined the output of the retrieval stage in a QA system by ensuring that semantic relationships between terms within a query were consistent with semantic relationships between the same terms in any returned documents.

Another aspect of our work is that we investigate whether sufficient information is present in individual document sentences to answer most queries unambiguously.

2. METHOD

For our analysis we relied on the TReC queries, the Aquaint [1] collection and Ken Litkowski's answer patterns and the lists of supporting documents.

This is the procedure we followed:

First all TReC factoid queries with non-nil answers, numbered between 2329 and 2393 were selected. There were 50 queries in total.

In every document answering one of the queries, we located the exact sentence containing the answer. We will refer to this as the supporting sentence. Supporting sentences from

duplicate documents were ignored.

Next each query was compared, term by term, with those in the sentence (or sentences if the answer occurred in more than one document) containing an answer. We recorded any instance where either a morphological (e.g. the same term, but in different inflections) or semantic relationship (e.g. synonyms) existed between two terms. In some cases it was necessary to chain these semantic relationships in order to link a query term to a term in a supporting sentence. For example "killer whale" can be linked to "orcas" as they are synonyms in different inflections.

If a query was not unambiguously answered by at least one supporting sentence it was noted. This might happen if the sentence depends on information from elsewhere in the supporting document. This judgement was made from the perspective of a person who was not an expert in the topic of the query.

Tables 1 and 2 list the types of morphological and semantic relationships encountered during our analysis. There are a couple of points to be noted from these tables:

- The Identical, DiffInf, DiffPOS, Synonym, Causal and WordChain relationship types are all treated as being symmetric. For these relationship types, it does not matter if one term is present in a query and the other is in the supporting sentence, or vice versa (e.g. "New York's" will always be a different inflection of "New York" regardless of which term is present in the query). The Hypernym, Hyperonym, Meronym, Holonym, UnitPair and CoOccur relationship types are all asymmetric (e.g. a *high* ⇒ *foot* UnitPair relationship is only considered valid if "high" occurs in the query.
- Our interpretation of what constitutes a different inflection of the same term is quite broad. It includes abbreviations and their expansions as well as variants of the same numerical or monetary value.
- We consider some terms to be so closely related that we have treated them as one. The result of this can be seen in the table's DiffPOS examples: France and French are really two distinct terms but we view them as merely being the same term in two different parts of speech.

The result of this process for one particular query is shown in Table 3.

3. RESULTS

Of the 50 queries, 44 were judged to have been answered unambiguously by at least one of their supporting sentences. The other six queries and their supporting sentences are shown in Table 4.

In four of these queries (i.e. 2338, 2379, 2382, 2388) this ambiguity can be attributed to the use co-references. Query 2382 in particular is interesting as the clause "The Bible itself contains two versions..." in the supporting sentence only implicitly refers to the topic of the query (i.e the Ten Commandments).

NYT19991217.0112	APW19980710.1204
car accident	car crash
Chapin wrote some strong story-songs, but he was still a work in progress when he died in a car accident in 1981.	1981 - Singer Harry Chapin killed in car crash in New York.
Chapin ⇒ *Identical* ⇒ *Chapin*	*Harry Chapin* ⇒ *Identical* ⇒ *Harry Chapin*
die ⇒ *Causal* ⇒ *car accident*	*die* ⇒ *Causal* ⇒ *car crash*

Table 3: Comparing Query Terms with Terms in Supporting Sentences for the unambiguously answered Query 2335, "How did Harry Chapin die?".

The supporting sentences for the other two queries (i.e. 2370 and 2376) require information relating to the context in which they occur. Potato chips, the topic of query 2370, are only mentioned in the sentences following the one listed in the table. Similarly the supporting sentence for query 2376 might be referring to a form of Martial Arts other than karate — we need more of the context to be sure.

There were 161 distinct supporting sentences for the 50 factoid queries we looked at. After removing prepositions, determiners and auxiliaries we found 439 instances where there was some morphological or semantic relationship between a query term and a term in one of its supporting sentences. That results in an average of 2.727 matches between a query and each of its supporting sentences. For the purpose of comparison, the average number of terms in each of the 50 queries, after eliminating prepositions, determiners and auxiliaries is 4.780.

We also determined the best supporting sentence for each query by selecting the one with terms morphologically or semantically related to the largest number of distinct query terms. Where a number of sentences were tied we favoured direct matches first followed by morphological relationships, followed by semantic relationships.

There were 166 query terms related to terms in the 50 best supporting sentences, which works out at 3.320 terms per query on average.

Table 5 shows the number of matches between query terms and terms in supporting sentences for each type of morphological relationship. However what is not shown in this table is the number of times it was necessary to chain relations together (e.g. *killerwhale* ⇒ *Synonym* ⇒ *orca* ⇒ *DiffInf* ⇒ *orcas*). There were 30 such instances (or 0.186 per supporting query on average) when all the supporting sentences were considered. There were nine chained relationships (or 0.180 per query) when we restricted our analysis to the best supporting sentences.

4. CONCLUSIONS

The objective of this paper was to assess the ways in which a query's terms could be linked to terms in its supporting sentences and to recognise any intrinsic performance limits

Relationship	Description	Examples	
Identical	The exact same term occurs in the query and in the supporting sentence.	Washington Monument	Washington Monument
DiffInf	The same term occurs in both the query and supporting sentence, but in different inflections.	New York	New York's
		$24.00	$24
		buy	bought
		MLB	Major League Baseball
DiffPOS	The same term occurs in both the query and supporting sentence, but in different Parts of Speech (POS).	France	French
		nominations	nominated
		hosted	host

Table 1: The types of morphological relationships examined by this paper. The relationship name, its description and some relevant examples are listed in the table. These relationship types are all symmetric.

Relationship	Description	Examples	
Synonym	A synonym of a query term is present in the supporting sentence.	killer whale	orca
		removed	deleted
		tall	high
Causal	There is a causal relationship between a query term and a term in the supporting sentence.	die	typhus
		have	lose
		invented	patented
		empty	flow
WordChain	A query term might be speculatively linked to a term in the supporting sentence via word chaining.	Oscar	best-actress
		musician	tracks
		hosted	Olympics
Hypernym	A query term is a hypernym of a term in the supporting sentence.	city	Berlin
		river	Hudson River
		instrument	cornet
Hyperonym	A query term is a hyperonym of a term in the supporting sentence.	navy	military
		Titanic	ship
Meronym	A query term is a meronym of a term in the supporting sentence.	Death Valley	California Desert
Holonym	A query term is a holonym of a term in the supporting sentence.	20th century	1945
		Death Valley	Greenland Ranch
UnitPair	For a given query term and a term in the supporting sentence, one is some physical attribute and the other is a unit that can be used to measure it.	high	foot
		hot	degrees
CoOccur	A query term might be speculatively linked to a term in the supporting sentence by frequent co-occurrence in queries and supporting sentences.	old	18
		many	696
		when	1910
		under	across

Table 2: The types of semantic relationships examined by this paper. The relationship name, its description and some relevant examples are listed in the table. The first three relationship types listed are symmetric and the remainder are asymmetric. In the latter types' examples column the query term is on the left and the document term is on the right.

Query No.	Query	Supporting Sentence
2338	When was the Titanic built?	The techniques used today to analyze the defects in the metal did not exist back in 1910 when the ship was being built, he said.
2370	When was the first potato chip made?	In 1853 railroad magnate Commodore Cornelius Vanderbilt complained that his potatoes were cut too thick and sent them back to the kitchen at a fashionable resort in Saratoga Springs, NY.
2376	What color belt is first in karate?	Every black belt, after all, started out as a white belt.
2379	What is the scientific name for red ants?	The ants are a particularly invasive species known scientifically as Solenopsis invicta, and they apparently reached California as hitch-hikers inside beehives shipped by a Texas beekeeper, according to James Brazzle, the University of California's entomology farm adviser for the Kern County cooperative extension service.
2382	What passage has the Ten Commandments?	The Bible itself contains two versions, one in Exodus 20:1-17 and a slightly different one in Deuteronomy 5:6-21
2388	What Arthur Miller play recounts his marriage to Marilyn Monroe?	Arthur Miller was criticized for "After the Fall," his 1964 two-act play featuring Maggie, a blonde, disturbed, childlike performer once married to the protagonist.

Table 4: Queries and Ambiguous Supporting Sentences.

	All Supporting Sentences		Best Supporting Sentences	
Relationship	Total	Per Sentence	Total	Per Sentence
Identical	215	1.335	81	1.620
DiffInf	59	0.366	23	0.460
DiffPOS	16	0.099	7	0.140
Morphological Total:	290	1.801	111	2.220
Synonym	11	0.068	6	0.120
Causal	16	0.099	6	0.120
WordChain	20	0.124	7	0.140
Hypernym	59	0.366	15	0.300
Hyperonym	13	0.081	5	0.100
Meronym	1	0.006	0	0.000
Holonym	4	0.025	2	0.040
UnitPair	9	0.056	1	0.020
CoOccur	16	0.099	13	0.26
Semantic Total:	149	0.925	55	1.100
Total:	439	2.727	166	3.320

Table 5: Frequency of Morphological and Semantic Relations. The table shows the frequency with which various relations are used in linking a query term to a term in a supporting sentence. These frequency counts are shown for all supporting sentences and for the best supporting sentences. In each case the average number of matches per supporting sentence is also provided.

associated with incorporating sentence level retrieval into a QA System.

One of the potential dangers of this type of the single sentence model is that information provided by supporting sentences might not be sufficient to determine a query's answer. It seems that this problem exists for a small minority of queries. That 44 out of 50 of the factoid queries in our investigation are answerable when only referring to the supporting sentence suggests a reasonably accurate QA system can be developed without recourse to other parts of the supporting document.

Capping the accuracy of a QA system at any level is still undesirable. Our results suggest that most ambiguous supporting answers require only co-reference resolution. However this observation is based on a very small sample size of six documents, where four required co-reference resolution to be useful in answering a query.

Our examination of the overlap which exists between query and document terms was illuminating. On average sentences with the answer to a query contain 1.335 of that query's terms. A well formulated query will increase this number to 1.620 for at least one supporting sentence. In reality most QA systems are able to search for a term in a number of different inflections. This could potentially increase the average number of query terms recognised in a supporting sentence to 2.080.

The goal of the retrieval stage of any QA system is to return documents (or sentences) containing an answer, while reducing the number other documents (or sentences) returned to a minimum. Relying on two query terms to locate answer-bearing sentences will frequently result in a large number of false positives among the retrieved sentences. However searching for all possible term matches using a range of morphological and semantic relationships is computationally prohibitive. Additionally implementing term matching for all the different relationships might simply not be worth the effort. Therefore we need to be selective about the relationship types we use to relax our query.

While the DiffPOS relation is relatively easy to implement it does not appear to provide much benefit as it is applicable on average to just 0.140 of supporting sentences. However it may be worth providing support for Hypernym and CoOccur relations as combined they are applicable to 0.560 of the best supporting sentences. They can also be implemented readily using predictive annotation. For example numbers within specific ranges could be identified in documents and tagged. This would help when identifying many CoOccur relations.

With sentence level retrieval, a QA system is substituting complexity in other parts of the system with what is intended to be a more effective retrieval stage. This study has revealed that under the most optimistic conditions the greatest number of query terms we can hope to match in a supporting sentence is 3.32. Given the increased importance of the retrieval stage in this type of QA system we need to raise this ceiling if at all possible. According to Riloff [5] prepositions can play a major role in text classification. Perhaps one way to improve the performance of a sentence level retrieval stage is to make more effective use of prepositions by including them in our search expressions.

5. REFERENCES

[1] Aquaint. Aquaint corpus. `http://wave.ldc.upenn.edu/Catalog/docs/LDC2002T31/`. Last accessed on 05/07/2004.

[2] C. Monz. Document retrieval in the context of question answering. In F. Sebastiani, editor, *Proceedings of the 25th European Conference on Information Retrieval Research (ECIR-03)*, volume 2633 of *Lecture Notes in Computer Science*, pages 571–579. Springer, 2003.

[3] M. Pasca. *High-Performance Open-Domain Question Answering from Large Text Collections*. PhD thesis, Southern Methodist University, 2001.

[4] J. Prager, D. Radev, E. Brown, A. Coden, and V. Samn. The use of predictive annotation for question answering in TReC 8. In *Proceedings of the Eight Text Retrieval Conference (TReC 1999)*, Gaithersburg, Maryland, 2000. National Institute of Standards and Technology.

[5] E. Riloff. Little words can make a big difference for text classification. In *Proceedings of the 18th International ACM SIGIR Conference on Research and Developement in Information Retrieval*, pages 130–136. Association for Computing Machinery, 1995.

[6] R. F. E. Sutcliffe, I. Gabbay, M. Mulcahy, and K. White. Question answering using the DLT system at TReC 2003. In *Proceedings of the 12th Text Retrieval Conference (TReC 2003)*, Gaithersburg, Maryland, 2003. National Institute of Standards and Technology.

[7] E. Voorhees and D. Tice. Building a question answering test collection. In *Proceedings of SIGIR 2000*, pages 184–191, Athens, Greece, 2000. Association for Computing Machinery.