Improving Complex Interactive Question Answering with Wikipedia Anchor Text

Ian MacKinnon[1] and Olga Vechtomova[2]

[1] David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada
imackinn@cs.uwaterloo.ca
[2] Department of Management Sciences
University of Waterloo
Waterloo, ON, Canada
ovechtom@engmail.uwaterloo.ca

Abstract. When the objective of an information retrieval task is to return a nugget rather than a document, query terms that exist in a document will often not be used in the most relevant information nugget in the document. In this paper, a new method of query expansion is proposed based on the Wikipedia link structure surrounding the most relevant articles selected automatically. Evaluated with the Nuggeteer automatic scoring software, an increase in the F-scores is found from the TREC Complex Interactive Question Answering task when integrating this expansion into an already high-performing baseline system.

1 Introduction

With the Complex Interactive Question Answering (CiQA) task introduced at TREC in 2006[1], the focus of evaluation is shifted from documents and facts to more elaborate nuggets. However, due to the concepts being sought having multiple terms to describe them, it becomes difficult to determine which sentences in the AQUAINT corpora of news articles contain the query terms being sought as they may be represented in the parent document by a variety of different phrases still making reference to the query term. For example, if the term "John McCain" was being sought, the phrase might appear in a document; however, the sentence which has the vital piece of information may simply contain "Senator McCain": an imperfect match.

In CiQA, templates are used with several bracketed items we call "facets" which are the basis for the information being sought. We can see from an example CiQA topic and answer key in Figure 1. A system must return text as a response which is then mapped to answer nuggets for scoring. Responses that correspond to 'vital' nuggets contribute to the score, 'okay' nuggets do not harm the score, and unassigned nuggets penalize the system for verbosity as a surrogate for precision[1].

Traditional query expansion of facets would introduce new terms which are related but do not necessarily mean the same as the original facet. This does not

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 438–445, 2008.

Qid 27: What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?

Topic	Number	Value	Nugget
27	1	vital	Mexico, Switzerland to cooperate on Salinas - Swiss seized over $114 million in bank accounts opened by Salinas
27	2	okay	Anti-drug police in Mexico confiscate 3.5 tons of marijuana
27	3	vital	Mexican heroin trafficking emerges - Mexican authorities discover a new organization smuggling heroin into the US
27	4	okay	Mexican navy seized 20 tons of cocaine off ships traveling Mexico's coast using technology and info supplied by American law enforcement
27	5	okay	Despite the often spectacular seizures and arrests, the bilateral structures to fight drugs put into effect by U.S. and Mexican governments... have been incapable of reducing the intensity of drug trafficking

Fig. 1. Templated query and answer key for a CiQA topic

always help the problem of query terms appearing in relevant documents but not within relevant sentences of the documents; it only introduces related terms which cannot be considered synonymous with the facet being retrieved.

Many of the CiQA facets are proper nouns and most thesauri, such as WordNet, do not contain entries for these. Thus, a new manner of finding synonyms must be found. In recent years, several new approaches have been proposed to use Wikipedia as a source of lexical information as it can be downloaded in its entirety and contains relatively high quality articles[2]. Wikipedia has previously been used in a lexical capacity to disambiguate named entities[3], explicitly compute semantic relatedness[4,5] and for word sense disambiguation[6].

As pointed out in previous work about creating an explicit semantic analysis engine based on Wikipedia[4], the anchor text which points to a Wikipedia article contains high quality terms which can be taken as synonyms for the articles which they link to. For example, the article "United States" will have frequent anchor texts such as "U.S.", "America", "American", "United States of America", or "USA".

While drawing potentially hundreds of articles becomes useful for semantic analysis, to find expansion terms we must first map facets to a small set of Wikipedia articles from which we can draw anchor texts to ascertain synonyms for the article title. Fortunately, by analyzing the whole Wikipedia corpus we can see the frequency of anchor text that links to articles.

We propose an algorithm to automatically select articles which best describe the facets of a CiQA topic in order to extract high quality phrases for expansion.

2 Wikipedia Article Selection for Facets

2.1 Automatic Article Selection Algorithm

We have devised a method of using the anchor text within Wikipedia links in order to resolve a small set of concepts which are represented in a candidate sentence.

Every article in Wikipedia represents a concept and all links from other articles to that article will have an anchor text associated with the link. We also know that there are Wikipedia guidelines for what the anchor text should be for a link, and that we can assume that, provided editors are following the rules, the anchor text of the link will be of high quality. As we can see from this excerpt from the Wikipedia manual of style[1]:

> "It is possible to link words that are not exactly the same as the linked article title, for example, [[English language—English]]. However, make sure that it is still clear what the link refers to without having to follow the link." -Wikipedia Manual of Style

The anchor texts which point to the article will contain other terms for the same concept which are necessary to get a better understanding of phrases that are used to describe the concept in the text. As we can see in Table 1, there are several different articles to the 'radio waves' anchor text of varying frequency.

Table 1. Frequency of links to articles that have "radio waves" as anchor text

Article Name	Anchor Text Frequency
radio waves	72
radio frequency	10
Electromagnetic radiation	3
radio	2
Radio Waves (album)	1

We define the algorithm to turn a facet into a list of concepts as follows:

1. Set window length to n.
2. For each possible position of window, check all anchor text in Wikipedia to see if the phrase or term is recognized. If it is, record the matching string and drop the words covered in the window from future consideration. See Fig 2.
3. Decrease the length of the window by one ($n = n - 1$). If the window length is 1, do not look up stopwords in term dictionary, simply ignore. Go to step 2 if window length is greater than 0.
4. For terms extracted from the query, look at the frequency of that term when linking to different articles. If an article has a majority of the links with that term as anchor text pointing to it, resolve that article to be the most relevant article for that multi-word unit. If no article has more than half the links with that anchor text pointing to it, drop the multi-word unit from consideration, as the term is ambiguous. However, if the frequency of anchor text linking to that article is less than 2, it is ignored.
5. If there are multiple articles resolved for the query, select whichever article has the highest number of incoming links from all other Wikipedia articles to be the most relevant Wikipedia article for the given facet.See Fig 3.

[1] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_%28links%29`

Fig. 2. When a window recognizes a multi-word unit from the nugget, it saves it and drops the text from future consideration

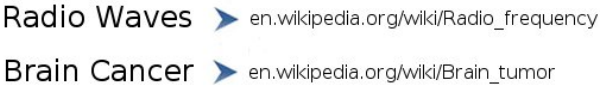

Fig. 3. Multi-Word Units are resolved to whichever article has the most links with that anchor text

In our experiments, we initially set $n = 5$.

By running this algorithm on the CiQA 2006 and CiQA 2007 test topics, we get sets of articles for every facet in each topic. To compare these automatically retrieved articles with the consensus articles of 12 human assessors, we use Fleiss' Kappa. Looking at this agreement, we find there to be a 0.6206 agreement between the human consensus articles and the automatically retrieved articles for the CiQA 2006 topics, and an agreement of 0.6764 for the CiQA 2007 topics. Both of these coefficients would be considered "substantial agreement" using the informal interpretation given by Landis and Koch[7].

We see a greater degree of agreement among the CiQA 2007 data, possibly on account of the more time-relevant data in the AQUAINT-2 corpus for Wikipedia. AQUAINT has articles from 1998 to 2000; before Wikipedia was launched. AQUAINT-2 has articles from when Wikipedia was considerably more popular, meaning the coverage of named entities from the news articles is likely more complete.

2.2 Baseline System

We base our system on the one which yielded the highest F-scores for initial automatic runs[8] at the CiQA 2006 task at TREC. To gain an initial set of documents, the system parses out the initial topic to get the 2 or 3 facets from the test topics, performs a BM25 retrieval[2] using the facet words as query terms, and returns the top 50 documents from the AQUAINT newswire corpus.

Once a list of documents has been retrieved, every document is split into candidate sentences. Preserving the rankings provided to us by BM25, we keep the sentences in order in which their parent document occurred in the top 50 ranking. Afterwards, a score of 0,1,2, or 3 to a sentence depending on the number of facets which are represented in a candidate sentence. Each topic will have 2 or 3 facets containing a number of terms within them. For each facet, let us consider $\Gamma = \gamma_1 ... \gamma_n$ to be the set of non-stopword terms for a facet in a CiQA topic.

[2] Using default parameters k=1.2, b=0.75.

A score is assigned to a candidate sentence S, by iterating through all the γ_i in Γ, and determining if any of the non-stopword stems of the terms exist in the sentence. If at least one exists, a nugget is said to be represented in the facet. More formally:

$$score(S, \Gamma) = \begin{cases} 1 & \text{if at least one of } \gamma_i \in \Gamma \text{ exist in S} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We get the total score for S by taking the sum of the $score(S, \Gamma)$ for each facet in the topic.

A sort is then performed on the list of sentences, but it is of great importance that the sort preserve the original ordering of the sentences with the same score. This allows for sentences which come from a document with a higher BM25 score to be ranked higher, given that they are likely more relevant to the test topic. The top $n = 30$ ranked sentences for each topic are output by the system.

3 Experiments

The only accurate way to judge a binary ("vital" or "okay") F-score for a CiQA run is to have human assessors assign system responses to answer key nuggets. However, this poses a problem for experimentation since the turnaround time for an assessment. For all experiments, the Nuggeteer system is used for determining the F-scores of the system responses as it has shown itself to have a highest correlation with human scores of all the automatic evaluation systems[9].

3.1 CiQA Runs

In order to test the ability of anchor text to improve ciQA retrieval, we must first introduce a method of using the articles we have selected for each facet to be integrated into the base CiQA system described earlier.

If, for a given facet Γ, we have corresponding Wikipedia articles which have anchor text linking to them, the set of anchor text phrases for that facet will be $A = \alpha_1\alpha_2...\alpha_n$, each α_i being an anchor text which links to one of the Wikipedia articles resolved for the facet, with a frequency across the Wikipedia corpus greater than 1. Ensuring that at least 2 articles link to the facet-corresponding one with the same anchor text will prevent potentially vandalized articles from introducing noise into the set of synonyms for the facet, A.

In the ideal situation, only one Wikipedia article is resolved for a facet, with not terms leftover from the facet. In this case, each α_i represents a high-quality phrase which multiple editors on Wikipedia have agreed is a reasonable referent for the concept being described in the linked article. Thus, we can use is it as a substitute for the facet being sought. However, we find that only 45 of the 72 facets, or 62.5%, of the CiQA 2006 facets fit this optimal case.

We modify the baseline system described earlier to incorporate the information from a facet's A set of anchor text in addition to the set of terms in the facet, Γ. A higher score is given to a candidate sentence, S, if it contains an anchor

text term from A in it as opposed to simply a term from the facet. More formally:

$$score(S, \Gamma, A) = \begin{cases} 1.2 & \text{if at least one of } \alpha_i \in A \text{ exist in S} \\ 1 & \text{if at least one of } \gamma_i \in \Gamma \text{ exist in S, and no } \alpha_i \in A \text{ exist in S} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The score of 1.2 is rather arbitrary. It just needed to be higher than 1, but low enough such that 2 matches from A would not be ranked higher than 3 from Γ. Experiments were conducted with various weighting techniques, but none garnered a significant change in scores.

Afterwards, sentences are sorted according to score as before. The only difference from the baseline system is the integration of the A terms from the anchor text. The remaining issue is what method is used to select the articles from Wikipedia for the given facet, for which we described an automatic method earlier.

To test this, we compare the baseline system F-score against the Wikipedia-enhanced system. The results of these runs can be seen in Table 2.

Table 2. F-Scores for CiQA runs using Nuggeteer

Run	F-score	Percent Improvement
2006 Baseline	0.3356	n/a
2007 Baseline	0.3388	n/a
2006 Wiki	0.3718	10.8%
2007 Wiki	0.3663	8.1%

From the table we can see a modest improvement in F-scores using the proposed Wikipedia method.

Looking at the individual results of the 30 2006 topics, we find that the automatic article selection improves F-scores in 8 of the topics, leaves 20 static (less than 2% change), and decreases 2.

Looking at the 2007 topics more closely, we see the most improved topics being "What evidence is there for transport of [automobiles] from [China] to [Russia]?" and "What effect does [glucosamine] have on [arthritis]?". The "automobiles" facet being expanded to include the term "car" being the probable cause for the former, and the expansion of "glucosamine" to include other marketed names for the drug for the later.

The most under-performing queries were "What evidence is there for transport of [illegal immigrants] from [Croatia] to [the European Union]?" and "What effect does [the Red Tide] have on [sea creatures]?". "The European Union" resolved to the political entity, thus the country names in the vital nugget were not contained within it. In this case, a "PART-OF" relation would need to be established. For example, "Italy" would be referenced in a vital nugget, so a system would have to recognize "Italy" as a potential substitute for "the European Union" since Italy is part of the European Union political entity; something that could be plausible by using category information. "Sea Creatures" resolved to the "Marine Biology"

article, which was fairly general and caused query drift. This is on account of a small number of links pointing to that article with that anchor text.

4 Conclusions and Future Work

We proposed an algorithm to automatically select a small set of relevant Wikipedia articles. This method was found to have a substantial amount of agreement with the consensus of the human assessors.

Using Nuggeteer, we were able to show a modest improvement in F-scores for CiQA topics which used the Wikipedia anchor text method of query expansion.

It is likely that a few well selected articles were enhancing the retrieval, which we selected in both cases, while the poorly selected ones were noise that was not affecting the retrieval.

This line of research introduces several new directions involving Wikipedia, which has shown itself to be an up and coming source for lexical information. The first being the resolution of articles from a query. We showed that many previous approaches looked at the selection of a large array of articles for traditional latent semantic analysis. However, our approach is close to ones involving WordNet, in that a small set of lexical data is sought. When trying to resolve an article for a given phrase, there are many interesting questions, such as disambiguation of the potentially multiple articles with similar titles and whether a term is significant enough to warrant resolving to an article. We hope to improve our article resolution algorithm by incorporating a part-of-speech tagger and word sense disambiguation tools to more accurately select articles.

Further work could also be done to fine-tune the procedure for extracting synonyms for articles by looking at anchor text. The current method of only taking anchor text which labels a link to an article with a frequency higher than 1 was mostly done because a lack of CiQA datasets meant that there could be no effective training set. Once more sets become available, statistical models could be found to give the most appropriate synonyms based on the distribution of the anchor text.

In the future we hope to also begin looking at a connectionist model of Wikipedia articles, treating every link in the corpus as a semantic link between two concepts. Clearly, weights on the links would depend on the strength of the semantic bond between two concepts. Using this method it may be possible to retrieve a list of high-quality related terms which could also be used to aid in nugget retrieval. More importantly, it could be used to find intersections of related terms between two facets.

References

1. Kelly, D., Lin, J.: Overview of the TREC 2006 ciQA task. SIGIR Forum 41(1), 107–116 (2007)
2. Giles, J.: Internet encyclopaedias go head to head. Nature 438(7070), 900–901 (2005)

3. Bunescu, R., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, April 2006, pp. 9–16 (2006)
4. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of The Twentieth International Joint Conference for Artificial Intelligence, Hyderabad, India (2007)
5. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence, Boston, Mass, July 2006, pp. 1419–1424 (2006)
6. Mihalcea, R.: Using Wikipedia for automatic word sense disambiguation. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Rochester, New York, Association for Computational Linguistics, April 2007, pp. 196–203 (2007)
7. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977)
8. Vechtomova, O., Karamuftuoglu, M.: Identifying relationships between entities in text for complex interactive question answering task. In: TREC (2006)
9. Marton, G., Radul, A.: Nuggeteer: Automatic nugget-based evaluation using descriptions and judgements. In: Proceedings of NAACL/HLT (2006)