Identifying Relationships Between Entities in Text for Complex Interactive Question Answering Task

Olga Vechtomova
University of Waterloo, Canada
ovechtom@uwaterloo.ca

Murat Karamuftuoglu
Bilkent University, Turkey
hmk@cs.bilkent.edu.tr

1 Introduction

In this paper we describe our participation in the Complex Interactive Question Answering (ciQA) task of the QA track. We investigated the use of lexical cohesive ties (called *lexical bonds*) between sentences containing different question entities in finding information about relationships between these entities. We also investigated the role of clarification forms in assisting the system in finding answers to complex questions. The rest of the paper is organised as follows: in section 2 we present our approach to calculating lexical bonds between sentences containing different entities, section 3 contains the detailed description of our systems, in section 4 we present the results, and section 5 contains discussions.

2 Relationships Between Entities in Text

We hypothesise that a lexical bond between two sentences can be used to predict whether they discuss the same topic. Since the task in ciQA is to find a piece of information (nugget) about a "relationship" between two entities, we propose to use lexical bonds in order to determine whether the contents of a sentence containing one entity is related to the contents of a sentence containing the second entity. In other words, by calculating lexical bonds we aim to find sentences that are likely to discuss something related to both entities.

In order to determine if two sentences have a lexical bond, we first need to identify lexical links that they have. According to Hoey (1991) a lexical link exists between two words in text that are related by means of one of the following: (1) simple lexical repetition, (2) complex lexical repetition, (3) simple partial paraphrase, (4) simple mutual paraphrase, (5) complex paraphrase, (6) superordinate, hyponymic and co-reference repetition. In the reported experiments we only considered lexical links formed by simple lexical repetition, i.e., a lexical link is considered to exist between two instances of the same lexeme, but with possible morphological variations, such as past tense forms of verbs, plural forms of nouns, etc. This is done by stemming the document representation in advance and calculating lexical links using stemmed terms[1] (excluding common stopwords). In an earlier work (Vechtomova et al., 2006) we experimented with lexical links formed by simple lexical repetition, synonymy and hyponymy (determined using WordNet), and found that repetition on its own performs as well as the other relationships in a document ranking task.

The number of lexical links (x) that must exist between two sentences in order for them to form a bond, was determined empirically in an earlier experiment, and x=1 was used in this study.

3 System description

3.1 Baseline runs

First, for each topic we build a query and retrieve n top documents using the Okapi retrieval system (Robertson et al., 1995). The query is built as follows:

We will refer to the bracketed text segments in the Template section of the topic as *facets*. All words from the facets are used as the query. If template id=2 ("What [relationship] exist between [entity] and [entity]?"), we do not extract the first facet ("relationship"). The only exception is when relationship type is "financial ties", in which case the first facet is represented by words "financial, money, funds, monetary". All proper nouns from the Narrative section, identified using Brill’s Part-of-Speech tagger (Brill, 1995), are added to the query as well. The resulting query is used to retrieve 200 documents using Okapi.

[1] Porter's weak stemmer was used for this purpose (Porter, 1980)

In the next step, we expand the facets as follows: if the facet contains an adjective derived from a proper noun, we identify its pertainym via WordNet and add it to the facet (e.g. adjective "English" and its pertainym "England").

Our next step is to get a subset of "valid" documents that contain at least one proper noun in each facet (if it has any). If none of the facets contain proper nouns, then the whole document set is considered "valid".

All "valid" documents are split into sentences. We experimented with two different sentence ranking methods. In the first one (UWATCIQA1) sentences are ranked as follows:

1. Number of facets the sentence contains. A sentence is considered to have a facet if at least one word from that facet is present. We discard all sentences that have no facets.
2. Resolve ties by the number of query terms the sentence contains.
3. Resolve ties by the number of lexical bonds the sentence has with the following sentences in the document. A sentence is said to have a lexical bond with another sentence if they have at least one lexeme in common.

In the second run (UWATCIQA2) the following ranking method was used:

1. Number of facets the sentence contains (as in UWATCIQA1).
2. Resolve ties by the number of query terms the sentence contains (as in UWATCIQA1).
3. Resolve ties by the average inverse document frequency (*idf*) of all nonstopwords in the nugget.

In order to exclude very long and verbose sentences we remove those whose length exceeds 50 nonstopword stems. We also remove (near-)duplicate sentences as follows: if a sentence has more than 50% non-stopword stems in common with any sentence ranked higher, remove it. Thirty top-ranked remaining sentences in each run are then submitted as nuggets to NIST.

3.2 Final runs

Our clarification form contained top 15 nuggets retrieved in UWATCIQA1 run. The evaluators were asked to select relevant nuggets. We used the feedback to clarification forms in two different ways as described below.

Run UWATCIQA3:
Nuggets which have lexical bonds with the nuggets selected by the users in the clarification form are included in this run. The ranking order of nuggets is the same as in UWATCIQA1 run.

Run UWATCIQA4:
Query expansion terms were extracted from the user-selected nuggets, ranked by Offer Weight (Robertson, 1990), and top 20 were added to the original query. Nuggets were extracted from the top 200 retrieved documents in the same way as for UWATCIQA1 run.

4 Results

The results of the runs submitted to ciQA are presented in Table 1.

Run	**F-Measure**	**MANuR**
UWATCIQA1	0.247	0.179
UWATCIQA2	0.246	0.172
UWATCIQA3	0.247	0.189
UWATCIQA4	0.268	0.186

Table 1: Official results of the runs submitted to ciQA

There is a statistically significant (t-test P<0.002) difference of 4% in MANuR between UWATCIQA1 and UWATCIQA2. This provides some evidence that the use of lexical bonds in nugget ranking helps retrieve relevant nuggets, compared to the use of average *idf* of terms in the nuggets. Difference in F-Measure between the two runs, however, was not significant.

Out of the total of 431 nuggets shown to users in the clarification forms in all topics, 274 (63.6%) were selected as relevant. Comparison of the two expansion runs shows that UWATCIQA4 is better than UWATCIQA3 in F-measure by 8% (not statistically significant). In MANuR, however, UWATCIQA3 is better than UWATCIQA4 by 1.4% (statistically significant).

All our runs perform better in F-measure than the median (0.209) calculated over 23 runs submitted by participants in the track. Two topics in UWATCIQA1 and three topics in UWATCIQA4 have the best F-measure.

5 Discussion

Having analysed the results of the two baseline runs, we noticed substantial inconsistency in nugget judgements by the evaluators. There were many cases where both runs A and B returned the same nugget, however it is only marked as containing the correct answer in run A, but not in run B, although none of the other nuggets in run B were marked to contain this answer. The judgement inconsistencies are summarised in Table 2.

Run	**Total nuggets retrieved**	**Answers found in the retrieved nuggets**	**Missed answers**
UWATCIQA1	838	138	28
UWATCIQA2	842	136	34

Table 2: Judgement inconsistencies observed in the two baseline runs

Such a large number of missed answers makes the evaluation rather unreliable, and prevents us from drawing any firm conclusions from the official results regarding the developed methods. One possibility would be to re-calculate the performance measures having restored the missing answers. This, however, is only a partial remedy. We think that a new methodology is needed to obtain more consistent judgements across systems.

Also, in the current evaluation methodology only the first nugget containing the correct answer is noted. When a nugget contains the correct answer but was ranked below the first nugget containing the same answer, the fact that it contains the correct answer is not recorded, which limits the usefulness of track judgements for further experiments. This could be avoided if evaluators note all nuggets that contain the correct answer, but only the first correct is still used in calculating the performance measures.

References

Brill, E. (1995) Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. Computational Linguistics, 21(4):543–565.

Hoey M. *Patterns of Lexis in Text.* Oxford University Press; 1991.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), pp. 130-137.

Robertson, S.E (1990). On term selection for query expansion. Journal of Documentation. 46(4), 359-364.

Robertson S. E., Walker S., Jones S., Hankock-Beaulieu M., Gatford M. (1995) Okapi at TREC-3. In Harman D. (Ed.) Proceedings of the Third Text Retrieval Conference, NIST, Gaithersburg, U.S., pp.109-126.

Vechtomova O., Karamuftuoglu M. and Robertson S. E. (2006) On Document Relevance and Lexical Cohesion between Query Terms. Information Processing and Management, 42(5), pp. 1230-1247.