

A little known fact is ... Answering *Other* questions using interest-markers

Majid Razmara and Leila Kosseim

CLaC laboratory
Department of Computer Science and Software Engineering
1400 de Maisonneuve Blvd. West
Montreal, Quebec, Canada H3G 1M8
m.razma@cse.concordia.ca, kosseim@cse.concordia.ca

Abstract. In this paper, we present an approach to answering “Other” questions using the notion of *interest marking terms*. “Other” questions have been introduced in the TREC-QA track to retrieve *other interesting facts* about a topic. To answer these types of questions, our system extracts from Wikipedia articles a list of *interest-marking* terms related to the topic and uses them to extract and score sentences from the document collection where the answer should be found. Sentences are then re-ranked using universal interest-markers that are not specific to the topic. The top sentences are then returned as possible answers. When using the 2004 TREC data for development and 2005 data for testing, the approach achieved an F-score of 0.265, placing it among the top systems.

1 Introduction

In this paper, we describe a method for answering a new type of questions: “Other”. Since 2004, the TREC Question Answering Track has introduced a new type of challenge: answering “Other” questions [1]. The test set consists of a series of questions relating to a particular target (or topic). Each question series consists of factoid questions, list questions and ends with exactly one “Other” question. For example, question series # 69 of TREC-2004 is:

69	Target: France wins World Cup in soccer
69.1	FACTOID When did France win the World Cup?
69.2	FACTOID Who did France beat for the World Cup?
69.3	FACTOID What was the final score?
69.4	FACTOID What was the nickname for the French team?
69.5	FACTOID At what stadium was the game played?
69.6	FACTOID Who was the coach of the French team?
69.7	LIST Name players on the French team.
69.8	OTHER Other

The answer to the “Other” question is meant to be interesting information about the target that is not covered by the preceding questions in the series, and should consist of a snippet of text, called a nugget, extracted from the AQUAINT document collection. To evaluate the answers, for each “Other” question, NIST assessors create a list of acceptable information nuggets about the target. Some of the nuggets are deemed *vital*, some are *okay* and others are *uninteresting*. Systems are then evaluated based on precision and recall of the nuggets, and ultimately the F-measure with $\beta = 3$ ¹. *Vital* and *okay* nuggets are evaluated differently: the number of *vital* nuggets are used to compute both recall and precision; while *okay* nuggets are used for precision only.

Answering “Other” questions is a difficult task because we don’t really know what we are looking for. There is no exact definition of what constitutes a *vital* and an *okay* answer and humans themselves may have different opinions about how interesting a nugget is. In fact, at TREC-2005, the University of Maryland submitted a manual run for the “Other” questions [2] where a human had identified manually what he considered to be interesting nuggets for each questions. This manual run was then submitted for judging along with automatic runs and received an $F(\beta = 3)$ score of 0.299. This low score seems to show that humans do not agree easily on what constitutes an interesting (*vital* or *okay*) piece of information.

The remainder of the paper is organized as follows. In Section 2 we discuss our approach in detail. Section 3 presents the results of the generated sentences with the 2004 and 2005 TREC data. Section 4 then presents related work, and finally in Section 5, we present future directions.

2 Answering “Other” Questions

Fundamentally, our approach to answering *Other* questions is based on the hypothesis that interesting sentences can be identified by:

1. target-specific interest marking terms (e.g. *Titanic* \Rightarrow *sinking*, *J.F. Kennedy* \Rightarrow *assassination*), and
2. universal interest marking terms (e.g. *first man on the moon*, *150 people died*)

To identify these interest marking terms, we did not use the AQUAINT document collection, where the answer should be found. The AQUAINT collection consists of newspaper articles that do not necessarily present the highlights of a target. An article presents detailed facts regarding the target but not an overview. A rich resource to find interesting facts related to many targets is an encyclopedia. Many target types are described and the content of each article is a short summary that highlights the most interesting facts – precisely what we are looking for. To find target-specific interest markers, we therefore used the Wikipedia online encyclopedia². Wikipedia contains more than 1 million

¹ which means that recall is three times more important than precision

² <http://en.wikipedia.org>

encyclopedic entries for various topics ranging from famous persons, to current events, to scientific information. The chances of finding an article on the topic of an *Other* question is therefore high, and we can extract potentially interesting terms from these entries without much noise. These terms are then searched in the AQUAINT document collection to extract interesting sentences that are then re-ranked using universal interest marking terms. Sentences with the highest scores are finally presented as interesting nuggets.

2.1 Finding the Wikipedia Article

The first stage to answering an *Other* question is to find the proper Wikipedia article. This process is shown in Figure 1. First, we generate a Google query using the target of the question. The target is first parsed, stop words are removed, and consecutive capitalized words are quoted together as a single term. Because verbs in the targets are usually in the present tense (e.g. “Russian submarine Kursk sinks”, “France wins World Cup in soccer”) while in the Wiki article, verbs are usually in the past tense (e.g. “It *sank* in the Barents Sea”, “The tournament was *won* by France”), they are not included in the query. The remaining words and quoted terms are then ANDed and sent to the Google API to search the Wikipedia sub-domain.

If several Wikipedia articles satisfy the query, the first one is taken. However, if no Wikipeage satisfies the query, then we try to loosen the query. Considering that quoted terms often have a non-compositional meaning, we keep them as is but OR single words. If this is not sufficient, then we gradually remove the last single word from the end. Finally, if still no Wikipedia article is found, then we simply drop Wikipedia and take the top N documents³ of the AQUAINT collection using the original query.

2.2 Extracting Target-Specific Interest Markers

After the Wikipage or top N AQUAINT documents are retrieved, interest-marking terms are extracted from the page (or pages). Because the Wikipedia entries consist of rather short documents (with an average of 400 words per article⁴), we only consider named entities as interesting terms. These are extracted with the GATE NE tagger⁵. If the number of terms of a specific semantic type (Date, Location, Person and Organization) is abnormally high (20 terms for each semantic type), then we assume that the page does not present a balanced overview of the highlights, but presents a specific point-of-view about the target and will therefore be biased towards that point-of-view. For example, if a Wikipedia article on an event (e.g. the *1998 World Cup*) contains a large number of person names, then we assume that the article is biased towards describing the people involved (e.g. the soccer players) as opposed to *Other* interesting information. To avoid

³ between 3 to 10 depending on whether the number of keywords is large enough

⁴ http://en.wikipedia.org/wiki/Wikipedia:Words_per_article

⁵ <http://gate.ac.uk>

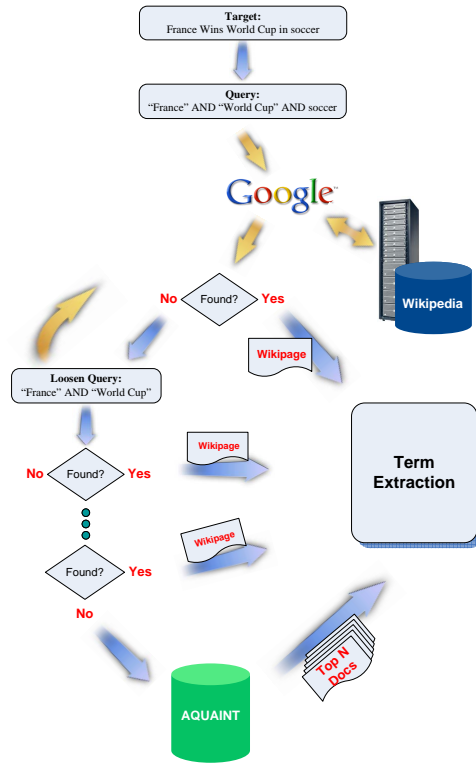


Fig. 1. Finding a Wikipedia article for the target “France wins World Cup in soccer”

this, we set a threshold on the number of terms for each semantic category that we keep. After removing terms occurring only once, the N most frequent terms are kept (in our case, 20).

We approximate co-reference resolution, by using word overlap. For example in answering the target “Port Arthur Massacre” we may find in the Wiki article the terms “Port Arthur” and “Port Arthur Massacre”. To consider both terms as a single concepts, we separate longer terms that overlap with shorter ones into sub-term (e.g. “Port Arthur” and “Massacre”).

2.3 Finding Interesting Sentences

Once we have a set of interesting terms for each target, we search for the N most relevant documents in AQUAINT. These documents are retrieved by the Lucene search engine⁶ using the same query generated for the target as in the Wikipage

⁶ <http://lucene.apache.org>

search (see section 2.1). If the appropriate Wikipage has been found then we also use a secondary query from the title of the Wikipage in order to get more documents related to the target. This secondary query is ORed to the Google query. For example, for the target “France wins World Cup in soccer” we have:

Google Query = “France” AND “World Cup” AND soccer
 Wikipage Title = 1998 FIFA World Cup

and we generate:

Lucene Query = (1998 AND “FIFA World Cup”) OR
 (“France” AND “World Cup” AND soccer)

If too many documents are returned through the Lucene search with this new query, then we add content words from the previous questions of that target (i.e. factoid and list question) to the query with less priority in order to focus the search. Since NIST also provides the output of the PRISE search engine with the target as query, we take the intersection of the top 25 documents returned by Lucene and the top 25 documents returned by PRISE. The idea is that if the two IR systems retrieved the same document using two different queries, then we should be more confident of its pertinence. Experimentally, we observed that taking the intersection of the two IR outputs increased the final F-measure by 0.02 with our testing set.

Within the documents chosen as the domain, the frequency of each interest marking term is then computed. For each term, we compute a weight as the logarithm of its frequency.

$$Weight(T_i) = Log(Frequency(T_i))$$

This weight represents how interesting a term is as a function of its frequency in the related documents. The less frequent a term, the less interesting it is considered.

2.4 Ranking Interesting Sentences

All sentences from the domain documents are then scored according to how interesting it is. This is computed as the sum of the weight of the interesting terms it contains.

$$Score(S_i) = \sum_{j=1}^n Weight(T_j) \quad | T_j \in S_i \quad \forall 1 \leq j \leq n$$

In order to increase the precision, we try to remove any extra characters on both ends of the sentence which do not contain much interesting material. Two kinds of information are removed: the source of the news at the beginning of sentences (e.g. *WASHINGTON (AP) - ...*) and markers of reported speech at the end of sentences (e.g. *..., local newspaper Daily Telegraph reported*).

After scoring the sentences and throwing away those with a score of zero (i.e. no interesting term in the sentence), we try to remove paraphrases. In order not to remove false paraphrases, we play it conservatively, and only remove lexically similar sentences. Either the sentences are almost equivalent to each other at the string level or they share similar words but not the same syntax. To compare sentences, we have used the *SecondString* package⁷, an open-source Java-based package of approximate string-matching techniques [3]. For removing the first kind of similarity, the Jaccard algorithm was used and for the second kind, the Jensen-Shannon was used. Both algorithms compute similarity based on token distance.

2.5 Universal Interest Markers

Once the sentences are ranked based on the target-specific interesting terms, we boost the score of sentences that contain terms that generally mark interesting information regardless of the topic. Such markers were determined empirically by analyzing the previous TREC data.

Superlatives We hypothesized that an interesting sentence would typically contain superlative adjectives and adverbs. People are interested in knowing about the *best*, the *first*, the *most wonderful*, and find normal or average facts uninteresting.

To verify this hypothesis, we computed the percentage of superlatives in *vital*, *okay* and *uninteresting* sentences from the 2004 data. For *vital* and *okay* sentences, we used the nuggets submitted by the 2004 participants and judged by the TREC assessors. For *uninteresting* sentences, we extracted sentences from the top 50 AQUAINT documents from the domain documents (see section 2.3) which do not contain *vital* or *okay* nuggets. The results, shown in Table 1, clearly show an increase in the use of superlatives in *vital* compared to *okay* and *uninteresting* sentences. When re-ranking nuggets, the score of a sentence that contains superlatives is therefore given a bonus. Experimentally, we set this bonus to be 20% of the original sentence score per superlative it contains.

Numerals We also hypothesized that sentences containing numbers probably contain interesting information also. For example, “Bollywood produces 800 to 900 films a year” or “Akira Kurosawa died at age 88”. To verify this, we also compared the percentage of numerals in *vital*, *okay* and *uninteresting* sentences on the same corpora. The results, shown in Table 1, again indicate that numerals are used more often in *vital* and *okay* sentences as opposed to *uninteresting* sentences. To account for this, the score of sentences containing numerals gets boosted by 20% for each numeral it contains. However, numerals that are part of a date expression such as *Sep 27, 2000* are excluded because we already considered them interesting terms from the Wikipedia entry.

⁷ <http://secondstring.sourceforge.net>

Sentence Type	Corpus Size	Superlatives	Numerals
<i>Vital</i>	49,102 words	0.52 %	2.46 %
<i>Okay</i>	56,729 words	0.44 %	2.26 %
<i>Uninteresting</i>	2,002,525 words	0.26 %	1.68 %

Table 1. Ratio of superlatives and numerals in each type of sentence

Interest Marking Keywords In addition to superlative and numerals, we also wondered if for specific target types, different terms are typically regarded as interesting. For example, information on someone’s birth or death, the founders of an organization, the establishment of an entity ... would all be considered interesting. These terms do not fit any specific grammatical category, but just happen to be more frequent in interesting nuggets. This is similar to the work of [4] (see section 4). To identify these terms, we analyzed the data of the 2004 *Other* questions. The data set consisted of:

1. The factoid and list questions of each target, because they mostly ask for interesting information.
2. The *vital* and *okay* answers to *Other* questions given by the TREC assessors⁸.
3. The actual answers to *Other* questions given by participants and judged *vital* and *okay* by NIST.

All these were stop-word removed and stemmed, then the frequency of each word was computed. The score of a keyword was computed as:

$$Score(K_i) = Freq(K_i) \times Distrib(K_i)^2$$

where $Freq(K_i)$ is the frequency of a keyword and $Distrib(K_i)$ is the number of targets whose sources contain the keyword. The intuition behind this scoring function is to favor keywords that are referred to in a high number of targets as opposed keyword that appears frequently, but only for a few targets. Hence a keyword K_i that occurs in a high number of targets is considered more important than a keyword K_j occurring more often (i.e. $Freq(K_j) > Freq(K_i)$) but in a smaller number of targets (i.e. $Distrib(K_j) < Distrib(K_i)$).

To identify terms that appear more often in interesting sentences as opposed to *uninteresting* sentences, we also built such a list of terms from the *uninteresting* answers submitted by the participants to the 2004 TREC QA (i.e. answers not considered as either *vital* or *okay*). Then, we computed the ratio of their scores as:

$$ScoreRatio(K_i) = \frac{Score_{int}(K_i)}{Score_{uni}(K_i)}$$

Where $Score_{int}(K_i)$ refers to the score of K_i in the *vital* and *okay* sentences and $Score_{uni}(K_i)$ refers to the score of K_i in the *uninteresting* sentences.

Table 2 shows the 15 top-ranking keywords that were extracted from all target types combined. As the table shows, the ranking of *most* and *first* verifies the importance of boosting superlatives.

⁸ available at http://trec.nist.gov/data/qa/2004_qadata/04.other_answers.txt

Rank	All Target Types	Thing	Person	Organization
1	found	kind	born	chang
2	die	fall	servic	publish
3	associ	public	serv	establish
4	life	found	become	first
5	begin	countri	film	leader
6	publish	offici	general	associ
7	first	field	old	larg
8	public	program	movi	found
9	servic	develop	chairman	releas
10	group	director	place	project
11	death	begin	receiv	group
12	see	discov	begin	lead
13	countri	particl	win	organ
14	old	power	life	begin
15	most	figur	intern	provid

Table 2. Interest-marking keywords in all target types and for each type of target

In order to make a specific list of interesting keywords for each target type, we did the same work for each category of questions (person, organization and thing). Table 2 also shows the list of frequent keywords per target type. Initially, we planned to consult a specific sublist according to the type of our target. For example, if the target is a person, then we only consult the person sublist. However, because we did not have much confidence in our target type tagger; we preferred to play it safe and we re-constructed a global list from the concatenation of the top 15 keywords of each sublist. This has two advantages to using the initial all-target type list. First, it allows us to make sure that each target type is equally represented in the global list. Although, the 2004 question set is not composed of thing, person and organizations targets in equal proportion, the 2005 question series contains equal number of questions for those target types. In addition, a re-constructed global list prevents us from considering terms that do not have a particularly high score in any one sublist, but occurs in every sublist with an average score; therefore having a high overall score, but not a high score in any one sublist (e.g. “see” and “most”). Sentences containing terms from the final re-constructed list are given a bonus of 20% per term, except if the term also appears in the previous questions of the target.

3 Results and Analysis

Once sentences have been extracted and sorted by their scores, they are evaluated. Since there exists no automatic standard scoring system for this task, we compared our sentences automatically to the assessor answers given by NIST and the actual answers submitted by all participants. If our sentence is identical to a *vital* or *okay* answer, we mark it as such. If our sentence is not identical

but is a substring of a longer *vital* or *okay* nugget, then to determine whether it contains the required information, we compare it to the assessor answers of that target (marked as *vital* or *okay*) using the token-based Jensen-Shannon similarity function⁹. If our sentence is closer to the assessor answer than the longer nugget is, then we consider our sentence as a correct one and mark it the same way the long answer is marked (*vital* or *okay*). Having a list of sentences marked as *vital*, *okay* or *uninteresting*, we can then evaluate the score of the question using the same F-measure (with $\beta = 3$) as used at TREC.

Since the TREC-“Other” task has only been introduced in 2004, we only have 140 such questions to develop and test the approach (65 questions for 2004 and 75 questions for 2005¹⁰). We therefore used the 2004 *Other* questions as the training set and the 2005 questions for testing. The results of the overall approach are shown in Table 3 along with the contribution of each type of universal marker. The figure marked *All* refers to the final score of the system when using all markers; while *All - X* refers to all markers except for X. *Best* and *Median* refer to the best and median score of all systems submitted to TREC-2005.

Markers Used	F-measure
All	0.265
All - Superlative Markers	0.255
All - Numeral Markers	0.257
All - Other Markers	0.266
Best	0.248
Median	0.156

Table 3. Test results with the 2005 *Other* questions

As the table shows, numeral and superlative markers increase the results somewhat; while, surprisingly, the keyword markers do not. We suspect that this is due to two main reasons:

1. To extract the interest marking keywords, a small corpus was used. We only had sentences related to 65 targets of 2004, which were composed of approximately 132,000 words; 44,000 words, for each of the three targets.
2. The TREC 2004 question series do not include the *event* target type; while this type of target accounts for 24% of the questions in 2005. Since we identified the keyword markers from the 2004 data, we have no specific markers for event types of target. In fact, if we compare the results of the approach per target type (i.e. Person, Event, Organization and Thing) we can clearly see that the F-score is lower for the *event* target type compared to the other target types (see Table 4).

⁹ <http://secondstring.sourceforge.net>

¹⁰ At the time of writing this paper, the TREC-2006 assessor judgments had not been released.

Target Type	Nb of Targets	F-measure
Person	19	0.300
Thing	19	0.277
Organization	19	0.268
Event	18	0.210

Table 4. Test results with the 2005 questions per target type

4 Related Work

Previous approaches to answering *Other* questions have mainly been addressed within the TREC confines, and only since 2004 [5, 6]. The most widely used approaches are based on patterns, keywords and question generation techniques.

In the pattern-based approach, a set of predefined patterns that seem to present interesting information are extracted from the answers of the previous years’ *Other* questions. Then the target is applied to the patterns to generate a potentially interesting string that is searched in the document collection. [7], for example, use a variety of strategies including the use of definition-patterns. For example, the pattern **TARGET, which...** is used to identify nuggets that define the target, and hence is deemed to contain interesting information. [8] also use patterns for extracting useful information and some semantic features to score sentences. These semantic features include comparative adjectives, digits, topic related verbs and topic phrases. [9] also use patterns and a summarizer based on lexical chains to extract a sentence as a summary of a passage.

On the other hand, keywords are also used to find the answers to *Other* questions. [10], for example, use syntactic information to identify interesting nuggets in the ACQUAINT collection. They identify sentences where the target appears in the subject or object position, then use a list of interest-marking keywords (similarly to our approach) to rank these sentences. [10] also uses the Wikipedia online encyclopedia to re-rank the sentences. However, they do not analyze the article per se to find interesting terms, but rather the corresponding XML file to look for the meta-data on the target and identify the categories the article belongs to. These categories are then used as keywords to re-rank the nuggets. As opposed to their work, we further re-rank the nuggets by using the universal interest markers. [11] identifies sentences that contain more than 50% of the words in the targets as candidate sentences. In ranking those sentences, those having more overlap with the target are given higher scores. Finally, [12] use statistics about word triplet co-occurrences from the documents related to each target then, extract snippets corresponding to the most frequent word triplets.

The third main approach used can be qualified as question generation that attempts to answer *Other* questions using *Factoid* or *List* question answering approaches. [7], for example, first classifies the targets according to their type, then creates a list of potential questions for each type of target. For example, if the target is of type *musician-person*, a set of questions such as *What is the name of the band of TARGET* or *What kind of singer is TARGET* are generated.

Using their factoid module, they then find answers for these typically interesting questions.

Some question answering systems use both pattern-based and keyword-based approaches. In [13], a web knowledge acquisition module determines which kind of knowledge base should be searched based on the target type. Then, the basic score of a candidate sentence is assigned either by searching the definitions about the target from online knowledge bases or by keywords and their frequencies. Finally, based on the target type, a set of structured patterns is used to re-rank the candidate sentences. [14] use a list of terms related to each target extracted from Web pages, Wikipedia and Britannica pages. Then Two types of patterns were used: lexical patterns (e.g. “X which is”, “like X”) and part-of-speech and named entity patterns (e.g. “TARGET, WD VBD”).

Other less popular approaches have also been proposed. For example, in [15], three strategies are exploited: a nugget can be extracted either by searching a database of definitional contexts, searching the corpus for a nugget including many keywords from the Websters Dictionary definition, or extracting all sentences from the top documents and using Wikipedia synonyms of the target. [7] also tries to locate specific named entities in the nuggets corresponding to the target types. For example, if the target is a person, then nuggets containing dates, quantities and locations are deemed more interesting.

5 Summary and Future Work

This paper proposed a keyword-based approach to extracting interesting sentences to answer *Other* questions. The method is based on the identification of target-specific and universal interest markers. Target-specific markers are identified by named entities found in the Wikipedia online encyclopedia. The frequency of these named entities in the AQUAINT documents are then used as a measure of how interesting they really are. Target-independent markers of interest are defined as the most frequent terms in the TREC-2005 *vital* and *okay* nuggets and include superlatives, numerals and specific keywords. Using these markers, we extract and rank sentences from the AQUAINT collection and return the top-scoring ones as the answer. When using the 2004 TREC data for development, the approach achieved an F-score of 0.265 with the 2005 TREC questions, placing it above the best scoring TREC-2005 system. We participated in TREC-2006, but the results have not been issued yet.

Currently, the system is highly dependent on the Wikipages; changing the term extraction source to something more robust (e.g. the top N web pages or top N AQUAINT documents) seems promising. In addition, we need to perform proper co-reference resolution on the Wikipedia terms; this would allow to better rank and identify the interesting terms. Also, computing lexical chains (as in [9]) may improve results as better target-specific markers can be identified; this needs to be investigated. Currently, to represent interesting facts, we only consider individual terms. A more precise method would ultimately be to expand the approach to extracting entire predicate structures; with roles and arguments.

Although the use of the universal keyword markers did not seem to improve results, we still believe it is an interesting venue. Since we have very little training data to identify these keywords, we plan to try to expand the ones we have with lexical semantics. Finally, since the result of event targets is rather weak, we need to focus more on this kind of targets.

Acknowledgement This research was financially supported by a grant from NSERC and Bell University Laboratories.

References

- [1] Voorhees, E.M.: Overview of the TREC 2004 Question Answering Track. [5]
- [2] Voorhees, E.M., Dang, H.T.: Overview of the TREC 2005 Question Answering Track. [6]
- [3] Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: Proceedings of the IJCAI Workshop on Information Integration on the Web (IIWeb), pages 73-78, Acapulco, Mexico (2003)
- [4] Ahn, K., Bos, J., Curran, J.R., Kor, D., Nissim, M., Webber, B.: Question Answering with QED at TREC-2005. [6]
- [5] Voorhees, E.M., Buckland, L.P., eds.: Proceedings of the Thirteenth Text REtrieval Conference, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2004)
- [6] Voorhees, E.M., Buckland, L.P., eds.: Proceedings of the Fourteenth Text REtrieval Conference, Gaithersburg, Maryland, National Institute of Standards and Technology (NIST) (2005)
- [7] Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A., Wang, P.: Employing Two Question Answering Systems in TREC-2005. [6]
- [8] Wu, M., Duan, M., Shaikh, S., Small, S., Strzalkowski, T.: ILQUA An IE-Driven Question Answering System. [6]
- [9] Ferres, D., Kanaan, S., Dominguez-Sal, D., Gonzalez, E., Ageno, A., Fuentes, M., Rodriguez, H., Surdeanu, M., Turmo, J.: TALP-UPC at TREC 2005: Experiments Using a Voting Scheme Among Three Heterogeneous QA Systems. [6]
- [10] Ahn, D., Fissaha, S., Jijkoun, V., Muller, K., Rijke, M., Sang, E.: Towards a Multi-Stream Question Answering-As-XML-Retrieval Strategy. [6]
- [11] Chen, J., Yu, P., Ge, H.: UNT 2005 TREC QA Participation: Using Lemur as IR Search Engine . [6]
- [12] Roussinov, D., Chau, M., Filatova, E., Robles-Flores, J.: Building on Redundancy: Factoid Question Answering, Robust Retrieval and the Other. [6]
- [13] Wu, L., Huang, X., Zhou, Y., Zhang, Z., Lin, F.: FDUQA on TREC2005 QA Track. [6]
- [14] Gaizauskas, R., Greenwood, M., Harkema, H., Hepple, M., Saggion, H., Sanka, A.: The University of Sheffield's TREC 2005 Q&A Experiments. [6]
- [15] Katz, B., Marton, G., Borchardt, G., Brownell, A., Felshin, S., Loreto, D., Louis-Rosenberg, J., Lu, B., Mora, F., Stiller, S., Uzuner, O., Wilcox, A.: External Knowledge Sources for Question Answering. [6]