TREC 2007 ciQA Task: University of Maryland

Nitin Madnani, Jimmy Lin, and Bonnie Dorr
University of Maryland
College Park, Maryland, USA
nmadnani,jimmylin,bonnie@umiacs.umd.edu

1 The ciQA Task

Information needs are complex, evolving, and difficult to express or capture (Taylor, 1962), a fact that is often overlooked by modern information retrieval systems. TREC, through the HARD track, has been attempting to introduce elements of interaction into large-scale evaluations in order to achieve high accuracy document retrieval (Allan, 2005). Previous research has shown that well-constructed clarification questions can yield a better understanding of users' information needs and thereby improve retrieval performance (Lin et al., 2006).

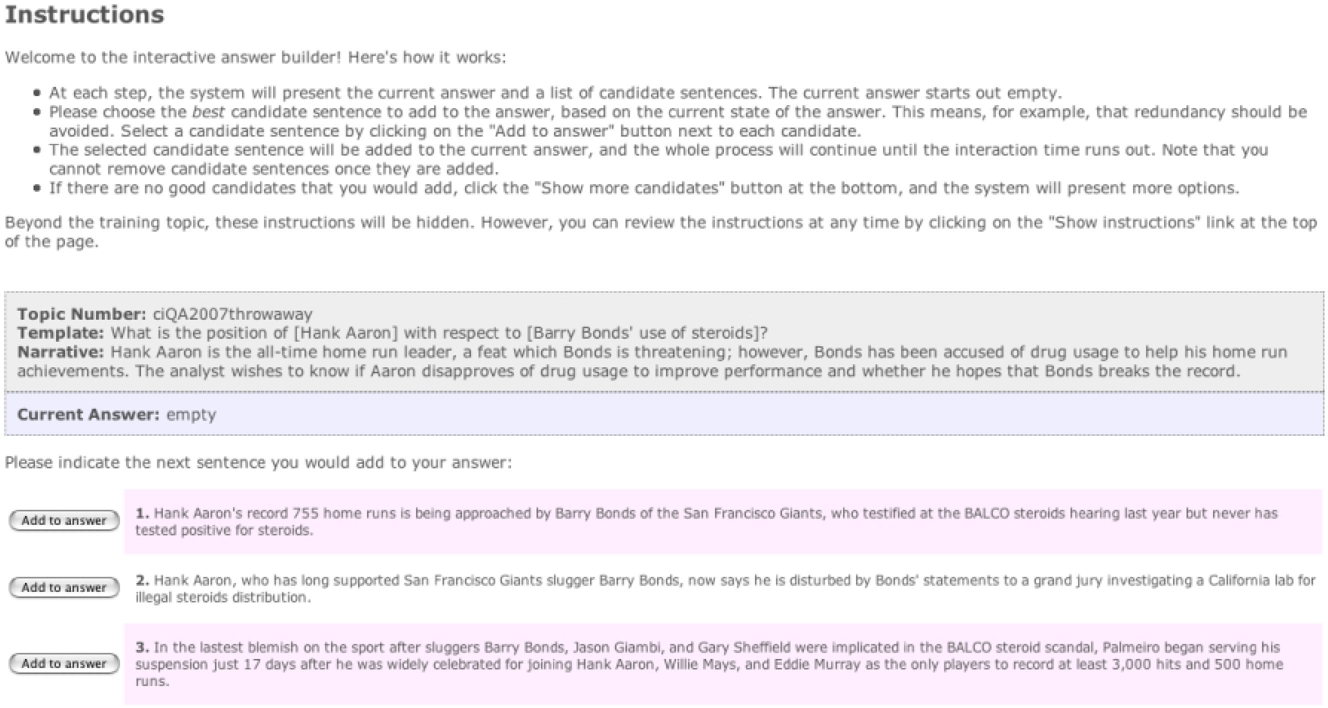
Interactive question answering has recently become a focus of research in the context of complex QA. The topics in the ciQA task are substantially different from factoid questions in that the information needs are complex, multi-faceted, and often not well defined or expressed. To investigate the role of interaction in complex QA, we experimented with two approaches. The first approach relied on Maximum Marginal Relevance (MMR) and is described in Section 2. The second approach employed the Multiple Alternative Sentence Compressions (MASC) framework (Zajic, 2007; Madnani et al., 2007), described in Section 3. Section 4 presents official results.

2 The MMR Approach

Maximum Marginal Relevance (MMR) is an extractive technique for document summarization (Goldstein et al., 2000). The basic idea is to iteratively select from a candidate pool of sentences until a desired length has been achieved. The score of a candidate sentence is computed from a weighted sum of two components: the relevance component and the redundancy component. At each step, the MMR algorithm selects the candidate with the highest score for inclusion in the final summary. Typically, the relevance component of a candidate sentence remains static, whereas the redundancy component is recomputed at each iteration as the summary grows in length.

2.1 Initial Run

In one set of experiments, we adapted MMR to ciQA. First, we retrieved the top 100 documents from the corpus with Lucene using the topic template verbatim as the query. These documents were then broken into individual sentences, which served as the pool of candidates. The relevance component of each sentence is computed as the sum of the *idf* of matching terms from the topic template. For the redundancy component, we used cosine similarity between each candidate sentence and the current answer. To answer a complex question, the sentence selector iteratively selects the candidate with the highest score, recomputing the redundancy component at each step. The process ends when 25 sentences have been selected in this manner. We submitted these answers to NIST as run `UMD07MMRa`.



Instructions

Welcome to the interactive answer builder! Here's how it works:

- At each step, the system will present the current answer and a list of candidate sentences. The current answer starts out empty.
- Please choose the *best* candidate sentence to add to the answer, based on the current state of the answer. This means, for example, that redundancy should be avoided. Select a candidate sentence by clicking on the "Add to answer" button next to each candidate.
- The selected candidate sentence will be added to the current answer, and the whole process will continue until the interaction time runs out. Note that you cannot remove candidate sentences once they are added.
- If there are no good candidates that you would add, click the "Show more candidates" button at the bottom, and the system will present more options.

Beyond the training topic, these instructions will be hidden. However, you can review the instructions at any time by clicking on the "Show instructions" link at the top of the page.

Topic Number: ciQA2007throwaway
Template: What is the position of [Hank Aaron] with respect to [Barry Bonds' use of steroids]?
Narrative: Hank Aaron is the all-time home run leader, a feat which Bonds is threatening; however, Bonds has been accused of drug usage to help his home run achievements. The analyst wishes to know if Aaron disapproves of drug usage to improve performance and whether he hopes that Bonds breaks the record.

Current Answer: empty

Please indicate the next sentence you would add to your answer:

1. Hank Aaron's record 755 home runs is being approached by Barry Bonds of the San Francisco Giants, who testified at the BALCO steroids hearing last year but never has tested positive for steroids.

2. Hank Aaron, who has long supported San Francisco Giants slugger Barry Bonds, now says he is disturbed by Bonds' statements to a grand jury investigating a California lab for illegal steroids distribution.

3. In the lastest blemish on the sport after sluggers Barry Bonds, Jason Giambi, and Gary Sheffield were implicated in the BALCO steroid scandal, Palmeiro began serving his suspension just 17 days after he was widely celebrated for joining Hank Aaron, Willie Mays, and Eddie Murray as the only players to record at least 3,000 hits and 500 home runs.

Figure 1: Interface for candidate sentence selection in the MMR approach. At each iteration, the user is asked to select a sentence for inclusion in the answer.

2.2 Interaction Design

At each step in the MMR algorithm, the sentence selector recomputes the score of all candidate sentences. Although only the highest-scoring candidate is selected for inclusion in the final answer, a ranked list of all candidate sentences is implicitly generated in the process. The idea behind our ciQA interaction involved this decision point: what if we presented the ranked list of sentences to a human and asked the human to select the best sentence for inclusion in the answer? Our MMR-based interactive QA system was exactly designed this way (see Figure 1). At each iteration, the user is asked to select from a ranked list of candidate sentences. The selected sentence is added to the final answer, and a new ranked list of sentences is computed based on the standard MMR algorithm. We asked the user to iterate this process for the entire duration of the interaction session. If the final output was still less than 4000 characters, we continued iterating the MMR algorithm (selecting the highest-scoring candidate) until that quota was filled. We submitted this run to NIST as `UMD07MMRb`.

3 The MASC approach

MASC (Multiple Alternative Sentence Compressions) is a framework originally developed for using sentence compression in query-focused automatic summarization of single and multiple documents. A MASC system uses a sentence compression module to generate multiple compressions of source sentences in combination with a candidate selector to construct a summary from the compressed candidates. The selector uses a combination of static and dynamic features to select candidates that maximize relevance while minimizing redundancy within the summary.

The MASC architecture consists of three stages: filtering, compression, and candidate selection. In the first stage (filtering), sentences of high relevance and centrality are selected for further processing downstream. In the second stage (sentence compression), multiple alternative compressed versions of the source sentences are generated, including a version with no compression, i.e., the original sentence. Our sentence compression module—Trimmer (Dorr et al., 2003; Zajic et al., 2006; Zajic, 2007)—uses linguistically-motivated trimming rules to remove constituents from a parse tree. It associates compression-specific feature values, such as the number and parse tree depth of various rule applications, with the candidate compressions that can be used in candidate selection. These features are computed and stored in advance for each compressed candidate. In addition, we also compute four features based on the query (topic template), the candidate compression, and the topic cluster (top n documents retrieved from the corpus with Lucene using the template verbatim as the query):

1. **Sentence Relevance**. The relevance score of the sentence to the query.
2. **Document Relevance**. The relevance score of the document to the query.
3. **Sentence Centrality**. The centrality score of the sentence to the topic cluster.
4. **Document Centrality**. The centrality score of the document to the topic cluster.

The final stage in MASC is the selection of candidates from the pool created by filtering and compression. We use a weighted linear combination of static and dynamic candidate features to select the highest scoring candidate for inclusion in the summary. Static features are those discussed above. Dynamic features include redundancy with respect to the current summary state and the number of candidates already in the summary from a candidate's source document. The dynamic features are recomputed after every candidate selection.

When a candidate is selected, all other candidates derived from the same source sentence are removed from the candidate pool. The selector continues to pick candidates for inclusion in the summary until either the summary reaches the prescribed word limit or the pool is exhausted.

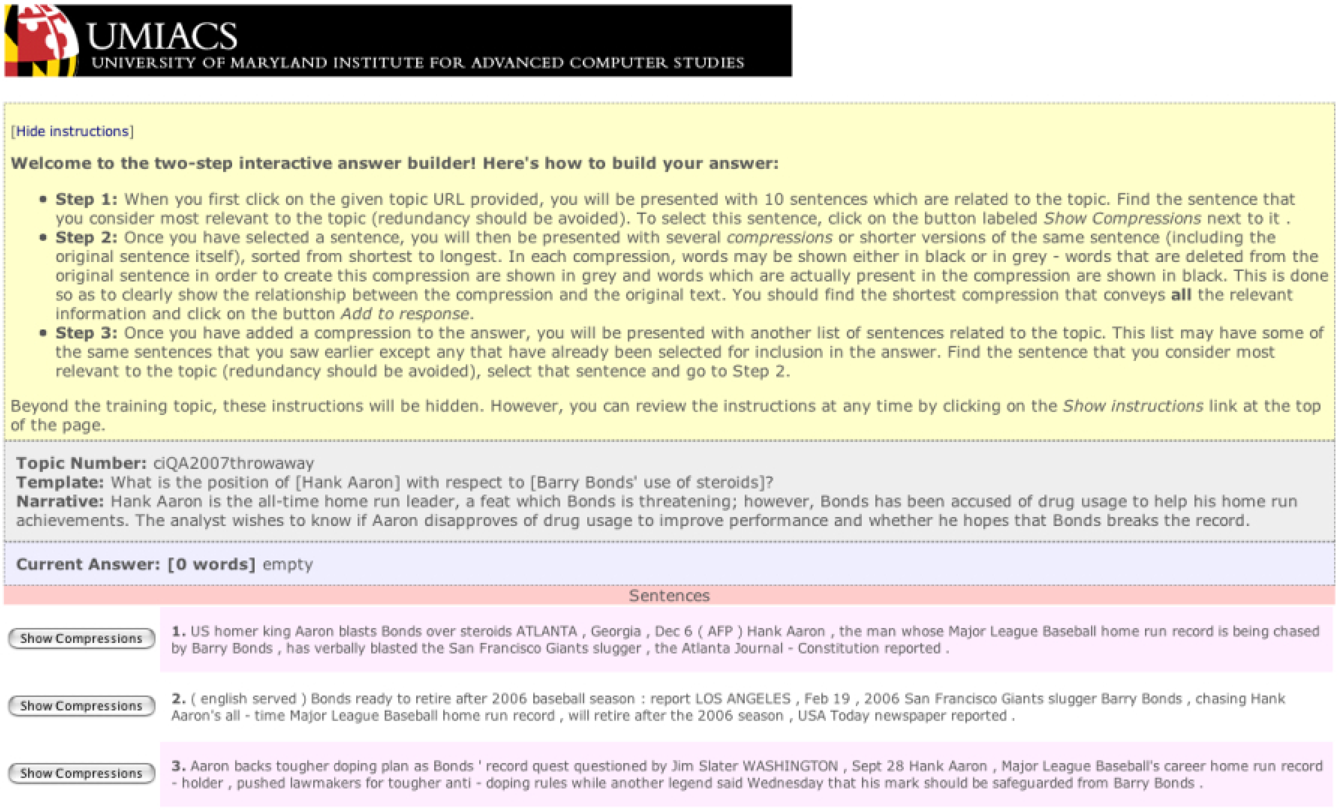
3.1 Initial Run

This section describes our methodology for creating the initial run `UMD07iMASCa` for the ciQA task. Answers were generated by treating this task as a query-focused multi-document summarization task. First we retrieved the top 50 documents for each topic from the corpus with Lucene using the topic template verbatim as the query. Next, we tagged and parsed all documents and ran Trimmer to generate compressions for all sentences. We merged the question template and the narrative for each topic into a single query that was used to compute the relevance and centrality features for each compressed candidate. We then ran our MASC system to completion with a word limit of 250 words. The weights for the candidate features were optimized on a separate held-out development set.

3.2 Interaction Design

This section describes the methodology used to create our final submission `UMD07iMASCb` based on the MASC approach. Since our original summarization system is completely automatic, we had to rearchitect it in order to incorporate the interaction with assessors. Processing up to the point prior to iterative selection of the answer candidates was the same as in the initial run.

Sentence Selection. We structured the system such that when first visiting the interaction URL, the assessors are presented with a list of the most relevant uncompressed sentences as determined by the Selector. After locating the most relevant sentence, the assessor is asked to click on the *Show Compressions* button next to the sentence. This redirects to another form that presents all compressions generated by Trimmer for that sentence (including the original sentence itself), sorted by length. See Figure 2 for an example.



Welcome to the two-step interactive answer builder! Here's how to build your answer:

- **Step 1:** When you first click on the given topic URL provided, you will be presented with 10 sentences which are related to the topic. Find the sentence that you consider most relevant to the topic (redundancy should be avoided). To select this sentence, click on the button labeled *Show Compressions* next to it .
- **Step 2:** Once you have selected a sentence, you will then be presented with several *compressions* or shorter versions of the same sentence (including the original sentence itself), sorted from shortest to longest. In each compression, words may be shown either in black or in grey - words that are deleted from the original sentence in order to create this compression are shown in grey and words which are actually present in the compression are shown in black. This is done so as to clearly show the relationship between the compression and the original text. You should find the shortest compression that conveys **all** the relevant information and click on the button *Add to response*.
- **Step 3:** Once you have added a compression to the answer, you will be presented with another list of sentences related to the topic. This list may have some of the same sentences that you saw earlier except any that have already been selected for inclusion in the answer. Find the sentence that you consider most relevant to the topic (redundancy should be avoided), select that sentence and go to Step 2.

Beyond the training topic, these instructions will be hidden. However, you can review the instructions at any time by clicking on the *Show instructions* link at the top of the page.

Topic Number: ciQA2007throwaway
Template: What is the position of [Hank Aaron] with respect to [Barry Bonds' use of steroids]?
Narrative: Hank Aaron is the all-time home run leader, a feat which Bonds is threatening; however, Bonds has been accused of drug usage to help his home run achievements. The analyst wishes to know if Aaron disapproves of drug usage to improve performance and whether he hopes that Bonds breaks the record.

Current Answer: [0 words] empty

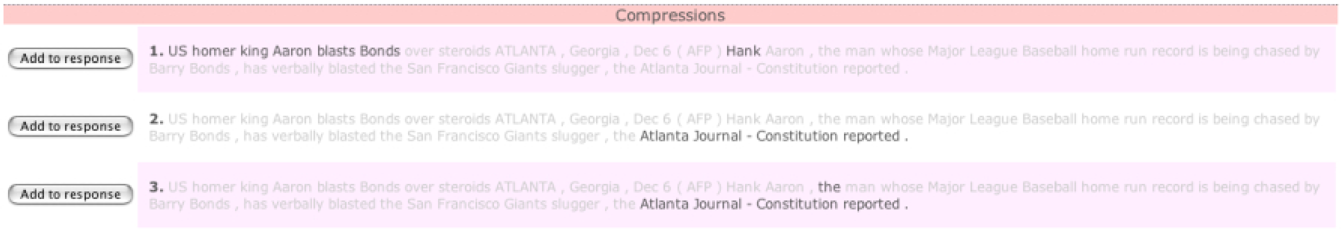
Sentences

1. US homer king Aaron blasts Bonds over steroids ATLANTA , Georgia , Dec 6 (AFP) Hank Aaron , the man whose Major League Baseball home run record is being chased by Barry Bonds , has verbally blasted the San Francisco Giants slugger , the Atlanta Journal - Constitution reported .

2. (english served) Bonds ready to retire after 2006 baseball season : report LOS ANGELES , Feb 19 , 2006 San Francisco Giants slugger Barry Bonds , chasing Hank Aaron's all - time Major League Baseball home run record , will retire after the 2006 season , USA Today newspaper reported .

3. Aaron backs tougher doping plan as Bonds ' record quest questioned by Jim Slater WASHINGTON , Sept 28 Hank Aaron , Major League Baseball's career home run record - holder , pushed lawmakers for tougher anti - doping rules while another legend said Wednesday that his mark should be safeguarded from Barry Bonds .

Figure 2: Interface for candidate sentence selection in the MASC approach. The user is first asked to select the most relevant sentence, and is then redirected to the interface shown in Figure 3 for selection of the best compression.



Compressions

1. US homer king Aaron blasts Bonds over steroids ATLANTA , Georgia , Dec 6 (AFP) Hank Aaron , the man whose Major League Baseball home run record is being chased by Barry Bonds , has verbally blasted the San Francisco Giants slugger , the Atlanta Journal - Constitution reported .

2. US homer king Aaron blasts Bonds over steroids ATLANTA , Georgia , Dec 6 (AFP) Hank Aaron , the man whose Major League Baseball home run record is being chased by Barry Bonds , has verbally blasted the San Francisco Giants slugger , the Atlanta Journal - Constitution reported .

3. US homer king Aaron blasts Bonds over steroids ATLANTA , Georgia , Dec 6 (AFP) Hank Aaron , the man whose Major League Baseball home run record is being chased by Barry Bonds , has verbally blasted the San Francisco Giants slugger , the Atlanta Journal - Constitution reported .

Figure 3: Interface for selecting the best compression of a candidate sentence in the MASC approach.

MMR		
type	runtag	pyramid F-score
initial	UMD07MMRa	0.333
final	UMD07MMRb	0.334

MASC		
type	runtag	pyramid F-score
initial	UMD07iMASCa	0.182
final	UMD07iMASCb	0.156

Table 1: Pyramid F-score of our submitted runs to the ciQA task.

Candidate Selection. Each compression in the list is color coded—words that have been deleted from the original sentence to create this compression are shown in grey and the words actually present are shown in black. From this list, the assessor is asked to find the *shortest* compression that conveys all relevant information and add it to the response. See Figure 3 for an example.

Once a compression has been added to the response, the assessor is presented with another list of sentences from the cluster. This list may have some of the same sentences seen earlier except for any that have already been included in the response. The process is repeated for as long as the assessor deems necessary (until the allotted interaction time runs out).

Automatic augmentation. Before submitting the final run, any answers created by assessors that are less than 250 words are further augmented with candidate selections made by our automatic summarization system until that word limit is reached.

4 Results

The pyramid F-scores of our submitted ciQA runs are shown in Table 1. It does not appear that our interactions were effective—the MMR interaction basically left the F-score unchanged, and the MASC interaction actually decreased the F-score.

References

James Allan. 2005. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

Bonnie J. Dorr, David M. Zajic, and Richard Schwartz. 2003. Hedge Trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC 2003)*, pages 1–8, Edmonton, Alberta.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Jamie Callan. 2000. Creating and evaluating multi-document sentence extract summaries. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM 2000)*.

Jimmy Lin, Philip Wu, Dina Demner-Fushman, and Eileen Abels. 2006. Exploring the limits of single-iteration clarification dialogs. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 469–476.

Nitin Madnani, David Zajic, Bonnie Dorr, Necip Fazil Ayan, and Jimmy Lin. 2007. Multiple alternative sentence compressions for automatic text summarization. In *Proceedings of the 2007 Document Understanding Conference (DUC-2007) at NLT/NAACL 2007*, New York.

Robert S. Taylor. 1962. The process of asking questions. *American Documentation*, 13(4):391–396.

David Zajic, Bonnie Dorr, Jimmy Lin, and Richard Schwartz. 2006. Sentence compression as a component of a multi-document summarization system. In *Proceedings of the 2006 Document Understanding Conference (DUC 2006) at NLT/NAACL 2006*, New York, New York.

David M. Zajic. 2007. *Multiple Alternative Sentence Compressions (MASC) as a Tool for Automatic Summarization Tasks*. Ph.D. thesis, University of Maryland, College Park.