

Putting the User in the Loop: Interactive Maximal Marginal Relevance for Query-Focused Summarization

Jimmy Lin, Nitin Madnani, and Bonnie J. Dorr

University of Maryland

College Park, MD 20742, USA

jimmylin@umd.edu, {nmadnani,bonnie}@umiacs.umd.edu

Abstract

This work represents an initial attempt to move beyond “single-shot” summarization to *interactive* summarization. We present an extension to the classic Maximal Marginal Relevance (MMR) algorithm that places a user “in the loop” to assist in candidate selection. Experiments in the complex interactive Question Answering (ciQA) task at TREC 2007 show that interactively-constructed responses are significantly higher in quality than automatically-generated ones. This novel algorithm provides a starting point for future work on interactive summarization.

1 Introduction

Document summarization, as captured in modern comparative evaluations such as TAC and DUC, is mostly conceived as a “one-shot” task. However, researchers have long known that information seeking is an iterative activity, which suggests that an interactive approach might be worth exploring.

This paper presents a simple extension of a well-known algorithm, Maximal Marginal Relevance (MMR) (Goldstein et al., 2000), that places the user in the loop. MMR is an iterative algorithm, where at each step a candidate extract c (e.g., a sentence) is assigned the following score:

$$\lambda \text{Rel}(q, c) - (1 - \lambda) \max_{s \in S} \text{Sim}(s, c)$$

The score consists of two components: the relevance of the candidate c with respect to the query q (Rel) and the similarity of the candidate c to each

extract s in the current summary S (Sim). The maximum score of these similarity comparisons is subtracted from the relevance score, subjected to a tuning parameter that controls the emphasis on relevance and anti-redundancy. Scores are recomputed after each step and the algorithm iterates until a stopping criterion has been met (e.g., length quota).

We propose a simple extension to MMR: at each step, we interactively ask the user to select the best sentence for inclusion in the summary. That is, instead of the system automatically selecting the candidate with the highest score, it presents the user with a ranked list of candidates for selection.

2 Complex, Interactive QA

One obstacle to assessing the effectiveness of interactive summarization algorithms is the lack of a suitable evaluation vehicle. Given the convergence of complex QA and summarization (particularly the query-focused variant) in recent years, we found an appropriate evaluation vehicle in the ciQA (complex, interactive Question Answering) task at TREC 2007 (Dang et al., 2007).

Information needs in the ciQA task, called topics, consist of two parts: the question template and the narrative. The question template is a stylized information need that has a fixed structure and free slots whose instantiation varies across different topics. The narrative is unstructured prose that elaborates on the information need. For the evaluation, NIST assessors developed 30 topics, grouped into five templates. See Figure 1 for an example.

Participants in the task were able to deploy fully-functional web-based QA systems, with which the

Template: What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?
Narrative: The analyst would like to know of efforts to curtail the transport of drugs from Mexico to the U.S. Specifically, the analyst would like to know of the success of the efforts by local or international authorities.

Figure 1: Example topic form the TREC 2007 ciQA task.

NIST assessors interacted (serving as surrogates for users). Upon receiving the topics, participants first submitted an initial run. During a pre-arranged period of time shortly thereafter, each assessor was given five minutes to interact with the participant’s system, live over the web. After this interaction period, participants submitted a final run, which had presumably gained the benefit of user interaction. By comparing initial and final runs, it was possible to quantify the effect of the interaction.

The target corpus was AQUAINT-2, which consists of around 970k documents totaling 2.5 GB. System responses consisted of multi-line answers and were evaluated using the “nugget” methodology with the “nugget pyramid” extension (Lin and Demner-Fushman, 2006).

3 Experiment Design

This section describes our experiments for the TREC 2007 ciQA task. In summary: the initial run was generated automatically using standard MMR. The web-based interactions consisted of iterations of interactive MMR, where the user selected the best candidate extract at each step. The final run consisted of the output of interactive MMR padded with automatically-generated output.

Sentence extracts were used as the basic response unit. For each topic, the top 100 documents were retrieved from the AQUAINT-2 collection with Lucene, using the topic template verbatim as the query. Neither the template structure nor the narrative text were exploited. All documents were then broken into individual sentences, which served as the pool of candidates. The relevance of each sentence was computed as the sum of the inverse document frequencies of matching terms from the topic template. Redundancy was computed as the cosine similarity between the current answer (consisting of

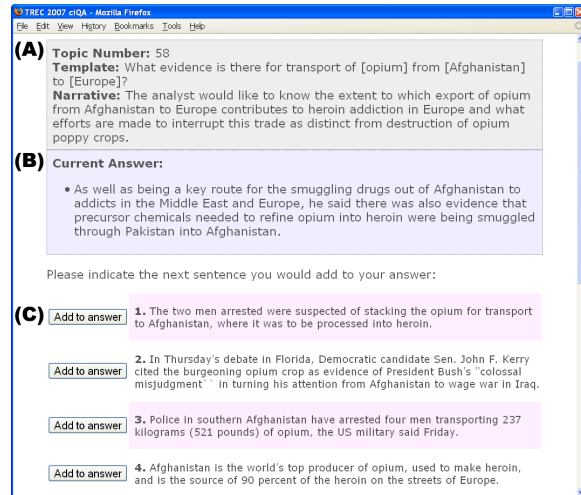


Figure 2: Screenshot of the interface for interactive MMR, which shows the current topic (A), the current answer (B), and a ranked list of document extracts (C).

all previously-selected sentences) and the current candidate. The relevance and redundancy scores were then normalized and combined ($\lambda = 0.8$). For the initial run, the MMR algorithm iterated until 25 candidates had been selected.

For interactive MMR, a screenshot of the web-based system is shown in Figure 2. The interface consists of three elements: at the top (label A) is the current topic; in the middle (label B) is the current answer, containing user selections from previous iterations; the bottom area (label C) shows a ranked list of candidate sentences ordered by MMR score. In each iteration, the user is asked to select one candidate by clicking the “Add to answer” button next to that candidate. The selected candidate is then added to the current answer. Ten answer candidates are shown per page. Clicking on a button labeled “Show more candidates” at the bottom of the page (not shown in the screenshot) displays the next ten candidates. In the ciQA 2007 evaluation, NIST assessors engaged with this interface for the entire allotted five minute interaction period. Note that this simple interface was designed only to assess the effectiveness of interactive MMR, and not intended to represent an actual interactive system.

To prevent users from seeing the same sentences repeatedly once a candidate selection has been recorded, we divide the scores of all candidate ranked higher than the selected candidate by two (an

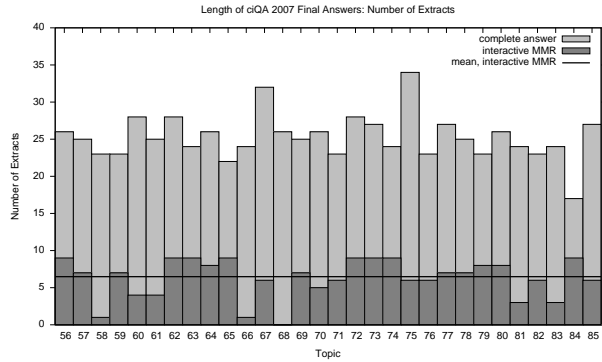


Figure 3: Per-topic lengths of final run in terms of number of extracts. Bars show contribution from interactive MMR (darker) and “padding” (lighter).

arbitrary constant). For example, if the user clicked on candidate five, scores for candidates one through four are cut in half. Previous studies have shown that users generally examine ranked lists in order, so the lack of a selection can be interpreted as negative feedback (Joachims et al., 2007).

The answers constructed interactively were submitted to NIST as the final (post-interaction) run. However, since these answers were significantly shorter than the initial run (given the short interaction period), the responses were “padded” by running additional iterations of automatic MMR until a length quota of 4000 characters had been achieved.

4 Results and Discussion

First, we present descriptive statistics of the final run submitted to NIST. Lengths of the answers on a per-topic basis are shown in Figure 3 in terms of number of extracts: darker bars show the number of manually-selected extracts for each topic during the five-minute interaction period (i.e., the number of interactive MMR iterations). The average across all topics was 6.5 iterations, shown by the horizontal line. The user selected the top ranking sentence only 28% of the time, and the average rank of the user selection was 4.9; the average length of answers (all user selections) was 1186 characters. Note that the interaction period included system processing as well as delays caused by network traffic. The number of extracts contained in the padding is shown by the lighter gray portions of the bars. For topic 68, the system did not record any user interactions (possibly resulting from a network glitch).

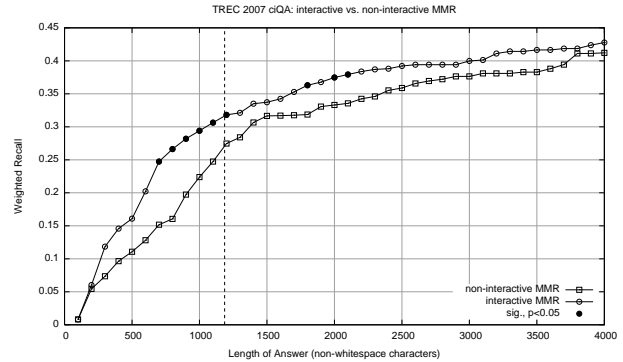


Figure 4: Weighted recall at different length increments, comparing interactive and non-interactive MMR.

The official metric for the ciQA task was F-measure, but a disadvantage of this single-point metric is that it doesn’t account for answers of varying lengths. An alternative proposed by Lin (2007) and used as the secondary metric in the evaluation is recall-by-length plots, which characterize weighted nugget recall at varying length increments. Weighted recall captures how much relevant information is contained in the system response (weighted by each nugget’s importance, with an upper bound of one). Responses that achieve higher nugget recall at shorter length increments are desirable in providing concise, informative answers.

Recall-by-length plots for both the initial run (non-interactive MMR) and final run (interactive MMR with padding) are shown in Figure 4, in length increments of 1000 characters. The vertical dotted line denotes the average length of interactive MMR answers (without padding). Taking length as a proxy for time, one natural interpretation of this plot is how quickly users are able to “learn” about their topic of interest under the two conditions.

We see that interactive MMR yields higher weighted recall at all length increments. The Wilcoxon signed-rank test was applied to assess the statistical significance of the differences in weighted recall at each length increment. Solid circles in the graph represent improvements that are statistically significant ($p < 0.05$). Furthermore, in the 700–1000 character range, weighted recall is significantly higher for interactive MMR at the 99% level.

Viewing weighted recall as a proxy for answer quality, interactive MMR yields responses that are significantly better than non-interactive MMR at

a range of length increments. This is an important finding, since effective interaction techniques that require little training and work well in limited-duration settings are quite elusive. Often, user input actually makes answers worse. Results from both ciQA 2006 and ciQA 2007 show that, overall, F-measure improved little between initial and final runs. Although it is widely accepted that user feedback can enhance interactive IR, effective interaction techniques to exploit this feedback are by no means obvious.

To better understand the characteristics of interactive MMR, it is helpful to compare our experiments with the ciQA task-wide baseline. As a reference for all participants, the organizers of the task submitted a pair of runs to help calibrate effectiveness. According to Dang et al. (2007), the first run was prepared by submitting the question template verbatim as a query to Lucene to retrieve the top 20 documents. These documents were then tokenized into individual sentences. Sentences that contained at least one non-stopword from the question were retained and returned as the initial run (up to a quota of 5,000 characters). Sentence order within each document and across the ranked list was preserved. The interaction associated with this run asked the assessor for relevance judgments on each of the sentences. Three options were given: “relevant”, “not relevant”, and “no opinion”. The final run was prepared by removing sentences judged not relevant.

Other evidence suggests that the task-wide sentence retrieval algorithm represents a strong baseline. Similar algorithms performed well in other complex QA tasks—in TREC 2003, a sentence retrieval variant beat all but one run on definition questions (Voorhees, 2003). The sentence retrieval baseline also performed well in ciQA 2006.

The MMR runs are compared to the task-wide reference runs in Figure 5: diamonds denote the sentence retrieval baseline and triangles mark the manual sentence selection final run. The manual sentence selection run outperforms the sentence retrieval baseline (as expected), but its weighted recall is still below that of interactive MMR across almost all length increments. The weighted recall of interactive MMR is significantly better at 1000 characters (at the 95% level), but nowhere else. So, the bottom line is: for limited-duration interactions, interactive

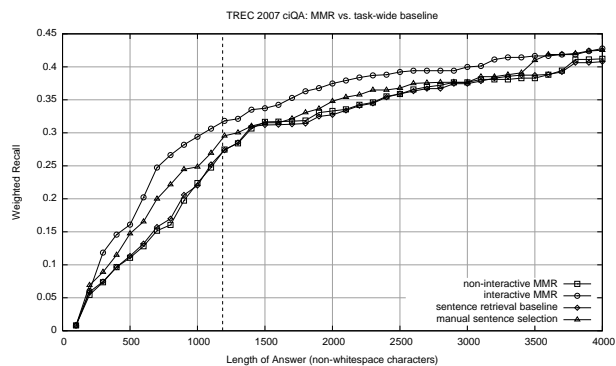


Figure 5: Weighted recall at different length increments, comparing MMR with the task-wide baseline.

MMR is more effective than simply asking for relevance judgments, but not significantly so.

5 Conclusion

We present an interactive extension of the Maximal Marginal Relevance algorithm for query-focused summarization. Results from the TREC 2007 ciQA task demonstrate it is a simple yet effective technique for involving users in interactively constructing responses to complex information needs. These results provide a starting point for future work in interactive summarization.

Acknowledgments

This work was supported in part by NLM/NIH. The first author would like to thank Esther and Kiri for their loving support.

References

- H. Dang, J. Lin, and D. Kelly. 2007. Overview of the TREC 2007 question answering track. *TREC 2007*.
- J. Goldstein, V. Mittal, J. Carbonell, and J. Callan. 2000. Creating and evaluating multi-document sentence extract summaries. *CIKM 2000*.
- T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *TOIS*, 25(2):1–27.
- J. Lin and D. Demner-Fushman. 2006. Will pyramids built of nuggets topple over? *HLT/NAACL 2006*.
- J. Lin. 2007. Is question answering better than information retrieval? Towards a task-based evaluation framework for question series. *HLT/NAACL 2007*.
- E. Voorhees. 2003. Overview of the TREC 2003 question answering track. *TREC 2003*.