Improving Passage Retrieval Using Interactive Elicition and Statistical Modeling

Daqing He[*]
Dina Demner-Fushman[†]
Douglas W. Oard[‡]
University of Maryland
College Park, Maryland 20742

Damianos Karakos[§]
Sanjeev Khudanpur[¶]
Johns Hopkins University
Baltimore, Maryland 21218

Abstract

The University of Maryland and Johns Hopkins University worked together in the 2004 High Accuracy Retrieval from Documents (HARD) track to explore design options for interactive passage retrieval systems. HARD assessors responded to clarification forms by (1) selecting additional search terms from an automatically constructed list of potentially discriminating terms, (2) selected relevant passages from an automatically constructed list of possibly relevant passages, and (3) entered additional search terms. Query expansion based on these three types of elicited information yielded statistically significant improvements in R-precision over baselines with and without blind relevance feedback. For topics that requested passages as answers, a preliminary analysis shows that statistical models for passage extent trained on HARD 2003 data yielded a significant improvement over a replication of the University of Maryland's HARD-2003 technique for passage extent determination, and the results of the new technique appear to generally be well above the median for HARD 2004 systems.

[*] Now at: School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260. Email: daqing@mail.sis.pitt.edu
[†] Email:demner@cs.umd.edu
[‡] Email:orar@glue.umd.edu
[§] Email: damianos@jhu.edu
[¶] Email: khudanpur@jhu.edu

1 Introduction

An Information Retrieval (IR) process can be modeled as establishing relationships between queries entered by a user and the documents in a collection. In such a model, evidence, usually counts of content-bearing words drawn from entire documents, is used as a basis for assessing the strength of a relationship. Some research has also focused on modeling portions of a document text (called "passage retrieval"), finding that passage-level evidence can sometimes provide better evidence for the full-document retrieval task than that of the full document text, especially when the documents are long or span different subject areas (Callan, 1994; Kaszkiel and Zobel, 1997).

We started to work on passage retrieval at the University of Maryland in the 2003 High Accuracy Retrieval of Documents (HARD) track. Our interest is motivated by the novel design of the HARD passage retrieval evaluation; in HARD passages are assessed based on their intrinsic utility to searchers as passages (rather than their extrinsic value as a basis for retrieval of full documents). For our 2003 experiments we developed a simple but effective module to identify and rank passages; it achieved an R-precision among the best reported that year. However, a subsequent inter-annotator consistency study conducted at the University of Maryland showed that our HARD-20003 passage retrieval module was far below human performance on the same task. Our analysis indicated that the most problematic part of our approach that year was passage extent determination; our passages were generally far shorter than the passages annotated by the HARD assessors. Therefore, the first research question we wanted to address was how we might better approximate hu-

man determination of passage extent. The Johns Hopkins University joined our team this year, and they focused on this challenge. We developed a set of paragraph-based features and some statistical models to identify the most likely passage extent for a query in its corresponding retrieved documents

Our second research question focused on optimizing the utility of a limited opportunity for user interaction. Previous research on presentation of passages in interactive information retrieval has focused on the display of passages as a basis for document selection task (Knaus et al., 1995; He et al., 2004). Our goal, by contrast, was to use passage-level feedback to improve passage retrieval effectiveness. We explored this by using passage selection and term selection in the design of our clarification forms, and then using the results as a basis for automatic query expansion.

In this report, we first introduce our 2003 passage retrieval module in section 2.1, then describe the design of the new passage extent model in section 2.2. We then move to a discussion of the design and use of clarification questions for improving passage retrieval in section 3. We conclude with a preliminary analysis of the experiment results in section 4.

2 Passage Retrieval

Perhaps the greatest challenge in the design of an end-user passage retrieval system is that there is little *a priori* basis for determining passage extent; some users may prefer terse passages, while others may prefer more context. Learning from examples can be useful in such cases, but only when users exhibit some degree of agreement regarding the desired passage length. In 2003, no training examples were available, however. We therefore adopted a simple ad-hoc approach for passage extent determination in HARD 2003. Since then, we have run a small inter-annotator agreement study, concluding that annotation consitency would be adequate to detect further improvements over our HARD 2003 system (Dina Demner Fushman et al., 2004). We therefore developed a new system for passage extent determination that is trained on the LDC HARD 2003 passage extent judgments. Both the old and the new system are described in this section.

2.1 The 2003 Passage Retrieval Module

Leveraging previous research on passage retrieval (Liu and Croft, 2002; Kaszkiel and Zobel, 1997), our 2003 passage retrieval module was based on assumptions that the relevance of a passage to a given query is related to:

- the computed overall probability of relevance for the document that contains the passage;
- the density of the query terms appearing in the passage;
- the importance of the query terms appearing in the passage.

For our 2003 passage retrieval module, we used (1) Inquery scores as a surrogate for the probability of relevance of the documents, (2) the number of different query terms appearing in the passage, and how close their positions in the passages are, as the representation of the query term density; and (3) TFIDF weights of the query terms in the passage, normalized for passage length, and adjusted by relative importance factors assigned to each query term based on its source (e.g., title field, clarification form, or blind relevance feedback) as the representation of the importance for each term. We formed a linear combination of these three factors, gaving more emphasis to the document scores because Inquery scores have been demonstrated to be a useful approximation to document relevance, whereas it was the first time we had tried the other two factors.

This approach implicitly assumes that passages have a known extent, but it offers no guidance on what that extent should be. We chose to model variable passage extents rather than using a fixed window size because that choice allows us to take advantage of paragraphs, a meaningful structural unit that is assigned by the author of the documents.

Our passage retrieval model identifies each instance of a query term and then extends the passage to the nearest paragraph boundary in each direction. When there is no paragraph markup in the document, we use a fixed 40-words windows size around the query term as the passage extent. When two passages containing a query term are adjacent, we merge them into a single larger passage.

Passages were ranked in decreasing order of these scores, and top 1000 passages were returned

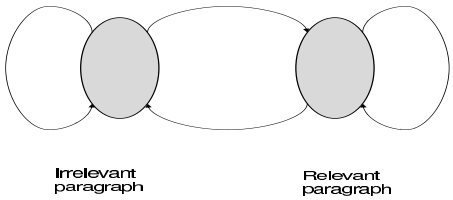

Figure 1: The Hidden Markov chain, with two states and four possible transitions. The output of the HMM is considered to be "emitted" either from a state or a transition.

for each query. In 2003, we limited the number of passages from a single document to the best three; based on our 2003 results, we have removed that restriction for 2004.

2.2 A Statistical Model for Passage Extent Determination

For 2004, we focused on improving passage retrieval by exploring (i) a novel Hidden Markov Model (HMM) approach, (ii) the application of Linear Discriminant Analysis (LDA), and (iii) a voting scheme among multiple classifiers.

Our analysis of HARD 2003 data showed that the atomic units for retreived passages are *paragraphs*; assessors worked on HARD 2003 data were never observed to choose sub-paragraph units as a passage. Therefore, we model each paragraph in a document as being in one of two states: "relevant" and "irrelevant". A "relevant" paragraph appears in the system output as part of a "relevant" passage.

It is natural to model the consecutive paragraphs in a document as changing their states according to the transitions of a Hidden Markov chain (see Figure 1). In this model, the probability that a paragraph is relevant or irrelevant *depends* on the characteristics of the current paragraph and the state of the immediately preceding paragraph. Table 1 shows the transition probabilities of the Markov chain, trained on a mixture of: (i) the HARD 2003 LDC passage retrieval relevance documents, and (ii) the top 100 passages that were found automatically by our 2003 passage retrieval system; these probabilities were used for the HARD 2004 evaluation.

	to relevant	to irrelevant
from relevant	0.87	0.13
from irrelevant	0.04	0.96

Table 1: The transition probabilities between "relevant" and "irrelevant" paragraphs.

We then assumed that the output of the HMM is a Gaussian-distributed scalar. Depending on the particular model, this scalar may be "emitted" at each state, or at each state *transition*. In both cases, this scalar-valued feature is equal to a linear combination of various similarity measures between the query and the paragraph, between adjacent paragraphs, etc. The weights of the linear combination can be determined, during the training phase, using Linear Discriminant Analysis (LDA), based on the ground truth.

We employed the following set of similarity measures for paragraph i as the scalar-valued features in our LDA model. All the paragraphs are preprocessed with stemming and stopword removal, and the temporal sequence of paragraphs in each document was preserved during the calculation.

1. *Paragraph features:* Similarity of the i-th paragraph with the query (a) Title, (b) Description, (c) Narrative, and (d) the "negative" portion of the query Narrative. (4 dimensions)[1]

2. *Document features:* Similarity of the entire document with the (a) Title, (b) Description, (c) Narrative, and (d) the "negative" portion of the Narrative. These provide a baseline for interpreting the similarity scores 1(a)-1(d) of individual paragraphs within the document. (4 dimensions)

3. *Document-minus-paragraph features:* Similarity of the document, less the i-th paragraph, with the (a) Title, (b) Description, (c) Narrative, and (d) the "negative" portion of the Narrative. (4 dimensions)

[1] By "negative," we mean the text segment in a narrative that describes what the retrieved passages "should not" contain. Such segments are found automatically by detecting the presence of cue phrases such as "should not contain."

4. *Inter-paragraph similarity:* Similarity between the i-th and the (i-1)-th paragraph. (1 dimension)

5. *Delta-features:* The "temporal" derivatives of the paragraph based features 1(a)-1(d), 3(a)-3(d) and 4. (9 dimensions)

All similarities are computed using the Okapi formula (Robertson et al., 1994), where the inverse "document" frequencies are computed at the paragraph level (i.e., they are inverse *paragraph* frequencies).

The elements of the above 22-dimensional vector are linearly combined through 3 sets of LDA coefficients: one set was trained assuming that the vector was emitted from the HMM state; the other 2 sets were trained assuming that the previous state was relevant or irrelevant, respectively (that is, the vector was emitted by the *transition* of the HMM).

During both training and testing, for each paragraph of each document, we computed scalar quantities equal to the linear combinations of the various similarity values and differences, with weights obtained through LDA training. Thus, we obtain 3 scalars: one assumed to be the output of an HMM with 2 conditional output distributions (one per state); and two scalars assumed to be the output of an HMM with 4 conditional output distributions (one per transition), where the two scalars correspond to two possible originating states.

For each one of the HMMs (one with outputs on states, and one with outputs on transitions), we compute the likelihood of the observed output using the forward-backward equations (Jelinek, 1997); then, we pick the state sequence which minimizes the *state* error (maximum aposteriori estimation).

Moreover, in addition to the HMM detectors, we used a collection of very conservative classifiers, which exploit some trends that were observed in the HARD 2003 data (and we assumed that these trends will also hold for HARD 2004). Specifically, we built the following 8 classifiers for finding relevant paragraphs in every retrieved document:

1. The paragraph with the highest similarity to the *title* field of the query is marked as relevant.

2. The paragraphs with the two highest differences of similarities (to the query's title) from the similarities of the preceding paragraphs are marked as relevant.

3. For each document, we expressed the similarities of paragraphs to their preceding paragraphs as a time series, and we computed its Fourier transform. Then, we set all paragraphs of a document to be relevant, if the bandwidth of the computed spectrum is among the lowest 10% bandwidths of all returned documents for a given topic. (By bandwidth we mean the range of frequencies which contains most of the signal energy.) The rationale behind this technique is that documents which are pretty homogeneous (i.e., all paragraphs are on-topic) have slow variation in inter-paragraph similarities (hence, small spectral bandwidth).

4-7. Similar to 1-2 above, but with similarity to the query's *description* and *narrative*.

8. The paragraph with the highest weighted sum of the values of its 22-dimensional vector (described above) is marked as relevant. The weights were chosen empirically, based on the HARD 2003 data.

Finally, each paragraph of each document is scored according to two schemes:

- **Score 1:** The number of classifiers which classify the paragraph as relevant (integer-valued). If no classifier mark it as relevant, and the adjacent paragraphs have non-zero scores, then Score 1 is equal to Score 2 (otherwise, if the adjacent paragraphs were not classified as relevant by any classifier, Score 1 is negative, and proportional to the "bandwidth" of the document).

- **Score 2:** The average of two normalized likelihoods of the paragraph, with respect to the two HMMs.

Since our passage retrieval model operates on the output of the document retrieval results, we trained on a mixture of documents: those which were truly relevant (obtained from the golden truth), and the documents which contained the top-100 passages that our document retrieval system had produced for the 2003 evaluation. We did a 10-fold cross-validation. Table 2 shows the R-precision obtained on HARD 2003 data, for different scoring schemes and test sets: (i) truly relevant

documents, obtained from the golden truth; (ii) The subset of UMD's (2003) retrieved documents that contained the top-100 passages for that topic; and (iii) top-1000 documents per topic. The measure reported throughout this report is R-Precision because this was the measure we used last year and during our training.

Score 1	
Testing on:	R-Precision
Truly relevant docs	0.51
Top-100 passages	0.37*
Top-1000 docs	0.23

Score 2	
Testing on:	R-Precision
Truly relevant docs	0.49
Top-100 passages	0.29
Top-1000 docs	0.12

Table 2: The retrieval effectiveness (R-Precision) obtained by JHU passage retrieval models on 2003 HARD data.

The 37% precision (marked with * above) is significantly higher than the 32% R-Precision that the UMD passage retrieval system achieved during the HARD 2003 evaluation.

Furthermore, one can see that, on average, Score 1 gives consistently better R-precision. For that reason, it was chosen as the first submission in the HARD 2004 evaluation.

During the development of the models, we also explored integrating Marti Hearst's TextTiling system (Hearst, 1997) into our passage retrieval model, where "tiles" were treated as atomic units rather than natural paragraphs. As shown in table 3, using 2003 passage retrieval data, the R-Precision under Score 1 was obtained for two TextTiling parameter values (w=7 and w=20), and following a 10-fold cross-validation procedure on the 1042 truly relevant documents and the top-1000 documents. In both cases, the R-Precision is significantly lower than the one obtained when the atomic blocks are paragraphs.

w	R-Precision	
	Truly relevant docs	Top-1000 docs
7	0.45	0.19
20	0.37%	0.11

Table 3: The Passage Retrieval Results of using "tiles" as atomic units

3 Clarification Questions

Communicating through clarification forms provides each site a means to interact with the people who proposed the search topic. We modeled the communication as a simplifed need negotiation process in our last year's HARD experiment (He and Demner-Fushman, 2003). Our work demonstrated that such interaction can be used to elicit several types of information, including relevance feedback, extra information about user's need, user's characteristics and user's preferences (He and Demner-Fushman, 2003). This year we mainly concentrated on just two of them – relevance feedback and extra information of the need, since they were found to be the most useful information last year.

As stated, the effectiveness of our passage retrieval module depends on the quality of the document returned, the query terms for finding the passage locations, and the extent of the passage. The passage extent problem was mainly addressed in section 2.2, however, we also took the chance of the interaction in clarification forms to ask user's performance of the passage length. One of the clarification question was

```
You expect your information need
to be fulfilled in/by:
  1. One or two sentences in a
     paragraph
  2. One or two paragraphs
  3. Several paragraphs in a
     document
  4. Several paragraphs in
     several documents
```

Eliciting named entities was demonstrated to be an effective approach for improving the search results in our last year experiment (He and Demner-Fushman, 2003), we, therefore, employed similar questions in this year clarification forms for name entities. We specifically worked on three types of named entities – personal names, organization names, and locations, all of which would give us phrases or other unique content words.

Our named entities related questions included relevance feedback questions, in which the terms were first identified by BBN IdentiFinder (bbn,), then selected based on the phrase's TFIDF scores from the top 10 returned documents. To satisfy the

space restrictions, we only selected top 5 ranked phrases for personal names, organization names, and locations respectively. The questions also included elicitation of extra terms in the same type.

The majority of clarification questions were dedicated to relevance feedback to returned passages if the users wanted passages as the preferred result format, or returned documents if otherwise. We knew from the topic metadata about the users' preference.

No matter which preference, we based the generation of clarification questions on the outcomes of our 2003 passage retrieval module. This was decided when we want to show the passages themselves if passages are to be judged since there would be some information lost no matter how good the summarization is, and our passage in general was short to be fit into the clarification forms. We identified that there is the screen space to up to five passages. If documents are to be judged, we are forced to use the surrogates, which are the concatenation of all the passages from those documents that are ranked within top 1000.

We had a choice of selecting top five ranked passages, or selecting top five passages from different sub-topic areas. We designed our clarification forms around the latter to let the user to view as many sub-topic areas as possible, and to maximize the possibility that some displayed passages are relevant even when the returned results were in poor quality. We designed a Maximum Marginal Relevance (MMR) like selection scheme to achieve the purpose.

Zhai defined MMR as a scheme that is capable of considering both the relevance and the novelty of returned documents (Zhai, 2002). Our MMR like selection scheme reflects this thinking. To maintain the relevance of the selected passages, we only chose passages that were ranked at top 200. Our selection of the number 200 was essentially ad-hoc, but it is a big number to include adequate number of different passages, and at the same time these passages are relative top ranked to maintain some relevance.

The novelty in our scheme was defined as the adequate difference between a passage and all previous selected passages. The difference was calculated based on the content terms in the passages. The weight of the terms was defined based on TFIDF. By starting the selection from the top ranked passages, our scheme identifies top five different passages.

We elicited two types of judgments from users. One type of judgments were related to the relevance of the passages/documents. The passage/documents could be "not relevant", "on topic" (i.e., soft relevant), or "relevant" (i.e., hard relevant). When the passages were displayed, the users were also asked to judge the length of the passages. Is the passage "too short", at the "right length", or "too long". We used the second type of information to fine tune the passage extent model for individual topics.

4 Experiments

4.1 Resources

The document retrieval system we used was InQuery text retrieval system (version 3.1p1) from the University of Massachusetts. The collection was the full HARD 04 collection, which contains 652,710 documents from eight different news sources. All the documents were stemmed using InQuery's own stemmer before indexing.

Before generating search questions, we preprocessed the topic statement. We marked up the named entities in the topic statement by using BBN's IdentiFinder, and treated them as phrases in queries. We also list the terms in the phrases as individual words in the queries for the case where only part of the phrases appearing in the documents.

4.2 Experiment Runs and Clarification Forms

We ran several baseline runs by using title only (run TITONL), title and description only (run TITDES), title plus phrases and top weighted (using TFAIDF) terms from description and narrative (run TFIDF), and the blind relevence feedback on top of the previous three runs (each is marked as run TOLBRF, TDABRF, and TIDBRF respectively in this report).

We generated two sets of clarification forms. CF1 was based on results from run TFAIDF plus utilizing the passage retrieval module to generating passages. CF2 was based on the BRF run of run TFAIDF (i.e., run TIDBRF), and it used the same passage retrieval module. We obtained users' answers for both sets.

Automatic query expansion was performed based on the answers from both CF1 and CF2 respectively. Highly representative content terms

were extracted based on TFIDF scheme from all the selected passages/documents for each topic. These terms combined with users provided NEs through clarification forms and the original queries became the expanded queries. The combination was weighted linearly with more weights to original queries and elicited NEs. Two document retrieval runs were generated based on the expanded queries obtained through this query expansion scheme, each of which corresponds to CF1 and CF2 respectively. They are marked as runs EXPCF1 and EXPCF2.

These document runs were then used as the input for both JHU passage retrieval models and our UMD passage retrieval models. Therefore, we generated three passage retrieval results run CF1JHU1, CF1JHU2, and CF1UMD, each of which corresponds to JHU model 1, JHU model 2 and UMD passage model.

5 Experiment Results and Discussion

5.1 Passage Results

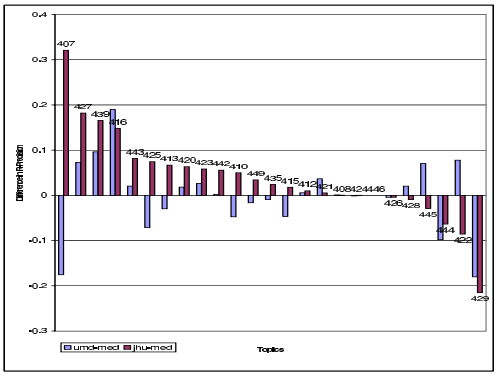

Figure 2: R-Precision difference between run JHUDOC1 and the medians of all the submitted runs on passage preferred topics.

As shown in Figure 2, both our runs CF1JHU1 (official run id UMAREXPR1) and CF1UMD (official run id UMAREXPR5) achieved reasonable well performance with most topics' results (measured by R-Precision) above the medians of all submitted runs. The run CF1JHU1, which uses JHU passage retrieval model Score 1, achieved 0.1468 average R-Precision. Our slightly improved UMD 2003 passage retrieval model achieved 0.0872 in average R-Precision. The difference between these two runs is 0.0586, and the difference is statistically significant, althought it is just under $P < 0.05$ using t-test.

To establish the potential of our passage retrieval model, we used the document golden truth as the input for our JHU passage retrieval model. As what we found in our training, the passage retrieval results improved dramatically. When marking passage extent in only relevant documents for the 2004 topics, the model based on Score 1 yields an R-Precision of 0.57, and that on Score 2 yields 0.55. This is an upper bound on passage extent performance with perfect document retrieval on Hard 2004 data.

5.2 Interactive Clarification

In our result analysis, we established two baselines for the comparison. The run TFAIDF mentioned above, which was used to generated CF1, does not have any feedback, so it was treated as a low baseline, whereas the blind relevance feedback run (run TIDBRF) is treated as a high baseline. The experimental runs are run CF1DOC, which is the expanded document run based on CF1, and run CF2DOC, which is the expanded document run based on CF2.

Our expanded runs achieved improvement over both baselines. Run CF1DOC obtained 21.20over the low baseline run TFAIDF, and the improvements is statistically significant (t-test $P < 0.05$). The improvement of run CF2DOC over the high baseline TIDBRF is 23.95(0.31 vs 0.2501), and the improvement is significant (t-test $P < 0.05$) too. The first improvement is similar to our last year's results (He and Demner-Fushman, 2003), but the second improvement is encouraging, it means that the clarification interaction can be combined with blind relevance feedback, and the improvement might be even bigger than performing interactive relevance feedback without blind relevance feedback first.

We then further explored the effectiveness of eliciting terms and relevance feeback on passage/documents seperately. As shown in Table 4, asking users to select relevant passages/documents yielded better improvement than elicting terms from users (0.2665 vs 0.2481 in CF1, 0.3188 vs 0.2671 in CF2), and the improvement achieved by the former runs over their corresponding runs that do not have clarification are statistically significant (t-test $P < 0.05$), whereas that of the latter runs

are not.

	R-Precision			
	no-exp	ask terms	feedback	all
CF1	0.2212	0.2481	0.2665	0.2681
CF2	0.2501	0.2671	0.3188	0.3166

Table 4: The effect of different approaches in interactive clarification measure by R-Precision.

6 Conclusion

In this report, we discussed our effort in exploring design options for interactive passage retrieval systems. We had two research questions to address: 1) how we might better approximate human determination of passage extent? and 2) how we could optimize the utility of a limited opportunity for user interaction? Our preliminary analysis of the results demonstrates that our newly developed passage retrieval model based on statistical modeling achieved significant improvement over our 2003 passage retrieval model, which was one of the best passage retrieval model in 2003. Our analysis also indicates that our design of interactions through clarification forms generated significant improvement over the baseline runs without the interaction, no matter whether or not the baseline employed blind relevance feedback. We also identified that aksing relevance feedback on documents/passages yield more improvement.

Our future work include further analyzing the experiment results, integrating users feedback on the passage length of individual topics into our passage retrieval model, and comparing the studies of interactive clarificaiton in both HARD 2003 and 2004.

Acknowledgements

Thank Jun Luo and Yejun Wu for their work on improving the UMD passage retrieval models and conducting experiments on TextTiling. Thank James Allan for organizing the HARD track, and folks at NIST and LDC for providing topics and relevance judgments. This work has been supported in part by DARPA cooperative agreement N66001-00-2-8910

References

Bbn identifinder. http://www.bbn.com/speech/docs/datasheets/idnt-022103.pdf.

Jamie Callan. 1994. Passage-level evidence in document retreival. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310.

Dina Demner Fushman, Daqing He, and Douglas W. Oard. 2004. Exploring Interactive Relevance Feedback With a Two-Pass Study Design. Technical Report LAMP-TR-116,CAR-TR-1001,CS-TR-4621,UMIACS-TR-2004-63, University of Maryland, College Park, October.

Daqing He and Dina Demner-Fushman. 2003. Hard experiment at maryland: From need negotiation to automated hard process. In *the proceedings of Text REtreival Conference (TREC) 2003*.

Daqing He, Jianqiang Wang, Jun Luo, and Douglas W. Oard. 2004. iCLEF 2004 at Maryland: Summarization Design for Interactive Cross-Language Question Answering. In *Proceeding of Cross Language Evaluation Forum (CLEF2004)*.

Marti A. Hearst. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Frederick Jelinek. 1997. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA.

M. Kaszkiel and J. Zobel. 1997. Passage Retrieval Revisited. In *In Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 178–185.

Daniel Knaus, Elke Mittendorf, Peter Schauble, and Paraic Sheridan. 1995. Highlighting Relevant Passages for Users of the Interactive SPIDER Retreival System. In *Proceeding of 1995 Text Retrieval Conference (TREC 4)*.

X. Liu and W. Croft. 2002. Passage retrieval based on language models. In *In Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382.

S. E. Robertson, S. Walker, Hancock-Beaulieu M., and M. Gatford. 1994. Okapi in trec3. In *Proceedings of Text Retrieval Conference TREC-3*, pages 109–126.

Chengxiang Zhai. 2002. *Risk Minimization and Language Modeling in Text Retrieval*. Ph.D. thesis, Carnegie Mellon University.