A Comparison of Sentence Retrieval Techniques

Niranjan Balasubramanian, James Allan and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
niranjan,allan,croft@cs.umass.edu

1. INTRODUCTION

Identifying redundant information in sentences is useful for several applications such as summarization, document provenance, detecting text reuse and novelty detection. The problem of identifying redundant information in sentences can be modeled as a sentence retrieval task. Given a sentence that contains some information, the task is to retrieve sentences in a given collection that express all or a subset of the same information.

Sentence retrieval techniques rank sentences based on some measure of their similarity to a query. The effectiveness of such a retrieval model depends on the similarity measure used to score sentences. An effective retrieval model should be able to handle low word overlap between the query and candidate sentences and go beyond just word overlap. Simple language modeling techniques like query likelihood retrieval have outperformed TF-IDF and word overlap based methods for ranking sentences. In this paper, we compare the performance of sentence retrieval using different language modeling techniques for the problem of identifying redundant information.

2. RELATED WORK

Previous work on novelty detection [9] and summarization[3] has explored several sentence level similarity measures and retrieval techniques for identifying novel information. Metzler et al [6] showed that query likelihood outperformed TF-IDF and word overlap based measures for identifying sentences that contain some specific facts contained in a query sentence. Murdock [8] compared query likelihood and a mono-lingual translation based model for the task of identifying restatements. Jeon et al[4] describe a mixture model of query likelihood and a translation based model for successfully identifying similar questions. Our work extends [6] and [8] by considering more sophisticated language modeling techniques such as topic based smoothing, dependence models and relevance modeling to improve retrieval effectiveness for identifying redundant information.

3. TEST COLLECTION

The experiments to detect redundant information were conducted on a dataset prepared by Murdock [8]. The dataset consists of documents and 50 queries that are answer sentences for some TREC 2003 QA track questions. Using query likelihood retrieval top 1000 documents(topic documents) were retrieved for each query. For each query, top 1000 retrieved documents were sentence segmented, stemmed using Krovetz stemmer and indexed. The queries were then used to retrieve sentences from their corresponding sentence indexes. For each query the top 50 retrieved sentences using three different retrieval models, query likelihood, Model-S [8], IBM Model-1 [1] were manually judged.

4. COMPARISON OF TECHNIQUES

We compare advanced techniques within the language modeling framework to improve sentence retrieval effectiveness.

4.1 Topic based Smoothing

Language modeling techniques primarily rely on estimates of word generation probabilities from document and query models. The estimates of word generation probabilities from small units of text such as sentences are not reliable and need to be smoothed. The word generation probabilities are smoothed using a linear interpolation of estimates from a topic model, built from the top 1000 documents retrieved from the document index for each query, and general English.

4.2 Dependence Models

Modeling query term dependencies has been shown to improve retrieval effectiveness for document retrieval [7]. We use the sequential dependence model described in [7] and ignore the full dependence model as it does not scale to long queries. Sequential dependence models capture term dependencies between consecutive terms in the query and relax the independence assumptions made by the query-likelihood model to some degree.

4.3 Relevance Models

Relevance modeling [5] is a technique for estimating query models from top ranked retrieved documents. Diaz et al [2] show that using a larger external corpus to build relevance models performs better than building relevance models using the target collection alone. We compare the effectiveness of relevance models built from target and external collections.

4.4 Translation based Models

Murdock [8] showed that Model-S, a translation based model, is effective for sentence retrieval, especially for question answering and novelty detection tasks. However, for the task of identifying redundant information at sentence level, Model-S is only as effective as a simple query likelihood retrieval. We compare the effectiveness of Model-S and a mixture model [4] using a large, relatively robust monolingual translation dictionary [4].

5. EXPERIMENTS AND RESULTS

Results are reported for the following retrieval techniques.

1. Query Likelihood Baseline (QL)
2. Topic Smoothing (QL-TS) - Collection estimates linearly interpolated with topic model estimates obtained from the top 1000 documents retrieved for each query.
3. Sequential Dependence Model (DM) - A weighted combination of the original query terms with a sequential dependence model query.
4. Translation Model (Model-S) - A translation based model described in [8] using a translation dictionary, Webfaq, built from FAQ pairs on the web [4].
5. Mixture Model (MM) - A mixture model of query likelihood and a translation based model described in [4].
6. Relevance Model-Target (RM-T) - Query model built using the target collection alone.
7. Relevance Model-External + Target (RM-E) - Query model built on a collection of external documents (Gigaword news corpus) and the target collection.
8. Interpolated Queries (RM+DM) - Best performing dependence model queries interpolated with best performing relevance model queries.
9. Two stage (DMRM) - Best performing dependence model queries used to retrieve documents that are then used to build relevance model queries.

Table 1: Comparison of effectiveness. *,** indicate significant ($p < 0.05$) improvements over QL-TS and DM, respectively, using a two-tailed paired t-test

Method	P@5	P@10	P@15	P@20	MAP
QL	0.6776	0.5531	0.4639	0.4102	0.6066
QL-TS	0.6857	0.5694	0.4735	0.4143	0.6248
DM	0.6980	0.5653	0.4735	0.4061	0.6264
Model-S	0.6735	0.5653	0.4803	0.4143	0.6189
MM	0.6735	0.5653	0.4748	0.4153	0.6198
RM-T	0.6939	0.5714	0.4789	0.4276**	0.6351*
RM-E	0.6980	0.5735	0.4912	0.4276**	0.6384*
RM+DM	0.7020	0.5755	0.4857	0.4133	0.6417*
DMRM	0.7061	0.5714	0.4789	0.4204**	0.6438**

Table 1 shows retrieval effectiveness in terms of precision at the top ranks and the mean average precision (MAP). Topic based smoothing and dependence models provide modest improvements over the query likelihood baseline. The translation based models provide no improvements over topic based smoothing. Using relevance models leads to small but significant gains over the best topic based smoothing run. Using a large external collection resulted in minor improvements over relevance models built from the smaller target collection alone. Dependence model queries provide a different form of evidence of relvance than relevance model queries and therefore their combination yields improvments over the individual queries. Also DMRM, which uses the best performing dependence model query to build relevance models, achieves the best performance by boosting the quality of the documents used to build the relevance models.

6. CONCLUSIONS

Previous work on sentence retrieval techniques show that simple query likelihood models outperform word overlap and TF-IDF based measures. In this short paper, we compared advanced language modeling techniques for the task of identifying redundant information in sentences and showed that they outperform simple query likelihood and topic based smoothing methods.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, and in part by NSF grant #IIS-0534383. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[2] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. *Proceedings of the ACM SIGIR conference*, pages 154–161, 2006.

[3] G. Erkan and D. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[4] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In *Proceedings of ACM CIKM*, pages 84–90, 2005.

[5] V. Lavrenko and W. Croft. Relevance based language models. *Proceedings of the ACM SIGIR conference*, pages 120–127, 2001.

[6] D. Metzler, Y. Bernstein, W. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. *Proceedings of CIKM*, pages 517–524, 2005.

[7] D. Metzler and W. B. Croft. A Markov Random Field model for term dependencies. In *Proceedings of the ACM SIGIR conference*, pages 472–479, 2005.

[8] V. Murdock. *Aspects of Sentence Retrieval*. PhD thesis, University of Massachusetts Amherst, 2006.

[9] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. *The Twelfth Text REtrieval Conference (TREC-12)*, 2004.