Brown at CL-SR'07: Retrieving Conversational Speech in English and Czech

Matthew Lease and Eugene Charniak
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI USA
{mlease,ec}@cs.brown.edu

Abstract

Brown's entry to the Cross-Language Speech Retrieval (CL-SR) track at the 2007 Cross Language Evaluation Forum (CLEF)[1] was based on the language model (LM) paradigm for retrieval [17]. For English, our system introduced two minor enhancements to the basic unigram: we extended Dirichlet smoothing (popular with unigram modeling) to bigrams, and we smoothed the collection LM to compensate for the small collection size. For Czech, time-constraints restricted us to using a basic unigram model, though we did apply Czech-specific stemming. While our English system performed well in the evaluation and showed the utility of our enhancements, several aspects of it were rushed and need to be addressed in future work. Our Czech system did not perform competitively but did provide us with a useful first experience in non-English retrieval.

1 Introduction

Our participation in the Cross-Language Speech Retrieval (CL-SR) track at the 2007 Cross Language Evaluation Forum (CLEF) represents our group's first effort in developing and applying an information retrieval (IR) system for human language. Our over-arching interest in this area is focused on two directions for future investigation [9]: (1) deeper syntactic/semantic analysis of queries and documents and (2) giving greater attention to speech-specific phenomena present in spontaneous speech. The system we developed for this year's CL-SR evaluation does not present new research in these areas but was intended to provide us with initial experience in retrieving conversational speech data and developing a competitive baseline model supporting future work.

Our retrieval system is based on the language model (LM) paradigm for retrieval in which a document's relevance is estimated as the probability of observing the query string as a random sample from the document's underlying LM [17, 25]. The unigram LM approach has been shown to have a strong theoretical connection to TF-IDF [26] and perform comparably [3] to other state-of-the-art approaches like vector similarity with pivot length normalization [22, 21] and the "probabilistic" approach [19, 24]. In addition to re-implementing the basic Dirichlet-smoothed unigram model, we also added two simple extensions. First, given the bag-of-words independence assumption underlying unigram and most other approaches to retrieval, modeling some notion of how words relate to one another seems like a useful step toward better modeling queries and documents. Previous work has shown that use of phrases yields around 10% relative improvement [1]. In the LM paradigm, use of phrases has been largely restricted to word pairs modeled via a unigram-bigram mixture model [23]. Whereas previous work has linearly mixed bigram and unigram models using a fixed mixture weight over all documents, subsequent work has

[1] http://www.clef-campaign.org

shown Dirichlet smoothing to outperform linear interpolation for unigram modeling by varying the mixture weight per-document according to document length. We extended document-collection unigram Dirichlet smoothing to include bigram mixing as well, as described in §3.

The second extension we added was collection smoothing. In the LM paradigm, the document LM is initially estimated by maximum likelihood, meaning any query word not observed in the document is assigned zero probability. This is problematic because there are likely many words related to the document yet which do not appear in it due to its brevity (i.e. chance). Assigning zero probability to a single term would zero-out the probability assigned to the entire query string, which is likely to be a poor estimate of document relevance. Consequently, it is common practice to smooth a document's LM with the collection LM (as a prior) to make the LM more robust. However, the collection LM is also estimated by maximum likelihood and so may also suffer from sparse data problems in the case of small collections. To investigate whether collection smoothing could help, we tried mixing the collection with larger corpora, and results show that collection smoothing substantially improved performance (§4).

Retrieval experiments were conducted in English and Czech (i.e. English queries/documents and Czech queries/documents), but due to time constraints we gave much less attention to Czech, evaluating only a unigram model without the above extensions. While our English system performed well in the evaluation, our Czech system was not competitive. §2 describes the data used in our experiments. In §3, system methodology is presented. Results are presented in §4, with additional details given in the appendix (§7). Concluding remarks appear in §5.

2 Data

This section introduces the English and Czech data used in our retrieval experiments. Our description is brief since this data has been previously described in detail [16, 15]. In addition to providing some factoid information about the collection, we also offer a few off-the-cuff reflections on our impressions of the broader significance of this task/dataset for IR.

The collection consists of interviews conducted by the Survivors of the Shoah Visual History Foundation (VHF) to record the memories of Holocaust survivors, rescuers, and witnesses. In terms of information management, cultural heritage archives like this one can be expected to become more frequent as recording and storage technology continues to become ever more widely accessible. Such archives are also just the tip of the iceberg in terms of the sorts of spontaneous speech being increasingly archived: debates, meetings, classroom discussion, talk shows, telephone conversations, online chat, etc.. In regard to information retrieval, previous work has considered broadcast news in detail [5] while spontaneous speech has garnered far less attention. Spontaneous speech also displays strikingly different phenomena than found in broadcast news or text with potentially interesting consequences for retrieval methodology. Word error rate is higher, topic segmentation is more problematic (potentially involving speaker identification and conversation untangling [2]), and indexing and use of retrieved content is complicated by back-channels, disfluency (filled pauses, explicit editing terms, self interruptions and corrections, etc.), and dramatically different sentential structure as speakers trail off, interrupt one another, and compose their utterances on-the-fly.

Compared to modern-sized retrieval collections, the VHF data set is quite small: the English collection consists of just 8,104 manually segmented interview passages to rank. However, whereas the size of the early text collections like Cranfield was limited by cost and human effort, a cultural heritage archive like this one may in fact be naturally small: in the case of VHF, there are only a limited number of people alive today with first-hand experience of the Holocaust. As for other forms of archived spontaneous speech, a particular individual may be interviewed only so many times, a course or talk shows series eventually terminates, etc. As such, we may increasingly see a practical need for effective search techniques on smaller collections such as this one; optimal methods and parameters on a terabyte collection may not yield the best performance here. Of course, one may have a broad information need and not care about which archived talkshow contained the relevant information. With regard to this scenario, the VHF collection does for spontaneous speech what early collections did for text, and likely larger collections of spontaneous speech are on the horizon.

Topics used were written in usual TREC style with three fields of increasing length: title, description, and narrative. These topics were based on actual information requests received by VHF from interested parties. Manual transcriptions of the interviews were not available in English or Czech, unfortunately, making it difficult to evaluate the impact of recognition errors on retrieval accuracy. Several variant one-best ASR transcripts were available for comparison. Both interviewer and interviewee were recorded on the same microphone/channel; while dialogue from the interviewee certainly dominates the interview, interviewer questions and comments are seen mixed into the same transcript.

2.1 English

Each segment contained a two to three sentence manual summary as well as a set of manually assigned keywords following a careful ontology developed by VHF. It is worth noting that these “manual” segments are far shorter than text and spoken documents commonly retrieved today; they are perhaps most akin to scientific paper abstracts in terms of previous retrieval experiments. Each interviewee also filled out a pre-interview questionnaire with some additional information that could also be used in the “manual” retrieval condition. The ontology developed by VHF could also be exploited for synonym expansion, etc. To date, the interview data and ontology have rarely been used [15]. For the “automatic” retrieval condition, two sets of automatically recognized keywords were available in addition to the ASR transcripts [15].

2.2 Czech

Czech interviews were not manually segmented as with English. Instead, track organizers supplied a set of scripts allowing the interviews to be automatically segmented for a fixed duration and overlap size with neighboring segments. They also supplied one such “quickstart” segmentation generated by their scripts, which had 11,377 segments of three-minute passages in which the first and last minute overlapped the neighboring segments. The idea of the scripts was to allow participants to explore the effects of various segmentations on retrieval accuracy. Unfortunately, problems with the scripts were found during the evaluation and led the organizers to suggest teams use the quickstart segments and avoid use of the scripts.

No manual summary or keywords available were available for the interviews. There also were no automatically recognized keywords; those used in the 2006 evaluation were removed for 2007 due to an unspecified problem with them.

3 Method

As described in the introduction, our retrieval system was based on the language model (LM) paradigm for retrieval [17]. In this paradigm, one assumes a unique language model (LM) underlies each observed document and estimates document relevance by the probability of observing the query as a random sample generated by the document’s underlying LM. Usually one assumes bag-of-words independence similar to that employed with the probabilistic and vector-space models: the probability of a string of words is computed as the product of the individual word probabilities (i.e. a unigram model).

$$P(Q|D) = \prod_{w \in Q} P(w|D) \tag{1}$$

One challenge of the LM paradigm is estimating the parameters of the underlying LMs given the brevity of the observed documents; if one simply takes the maximum likelihood estimate (MLE), a single query term unobserved in the document would zero-out the entire probability of observing the query given the document, making the entire framework exceedingly fragile. Instead one commonly employs smoothing to discount the probability mass assigned to observed terms and reserve some probability mass for all unseen terms. Previous work has shown Dirichlet smoothing of the form below to be most effective in practice [26]. This form of smoothing adds a hyper-parameter number of pseudo-counts distributed fractionally according to the prior collection model. N is the total number of words in the document.

$$P(w|D) = \frac{C_D(w) + \mu P(w|C)}{N + \mu} \tag{2}$$

The LM approach has been shown to have a strong theoretical connection to TF-IDF [26] and perform comparably to vector similarity and probabilistic approaches in practice [3]. A potential advantage of the LM approach lies in the pre-existing theoretical foundation and set of proven estimation techniques developed by earlier work in speech recognition.

A well-known improvement to the basic unigram LM is to model word pairs in a unigram-bigram mixture model [23]. While previous work linearly mixed bigram and unigram models using a fixed mixture weight over all documents, this misses the key idea of Dirichlet smoothing that longer documents provide more evidence for MLE and so should require less smoothing. For this reason, we extended document-collection unigram Dirichlet smoothing equation above to include bigram mixing. The Dirichlet smoothed bigram is given by:

$$P(w_i|w_{i-1}, D) = \frac{C_D(w_{i-1}, w_i) + \mu_1 P(w_i|w_{i-1}, C)}{C_D(w_{i-1}) + \mu_1} \tag{3}$$

and can be mixed with the unigram by adding an additional hyper-parameter, μ_2.

$$P(w_i|w_{i-1}, D) = \frac{C_D(w_{i-1}, w_i) + \mu_1 P(w_i|w_{i-1}, C) + \mu_2 P(w|D)}{C_D(w_{i-1}) + \mu_1 + \mu_2} \tag{4}$$

This leaves three hyper-parameters for tuning: μ, μ_1, and μ_2.

Our second extension to the basic LM approach was to incorporate collection smoothing. While it is common practice to smooth a document's LM with the collection LM (as a prior) to make the LM more robust, as shown above, the collection LM is also estimated by maximum likelihood and so may also suffer from sparse data problems in the case of small collections. To investigate whether collections smoothing could help, we tried linearly mixing the collection with two larger text corpora: 40,000 sentences from the Wall Street Journal as found in the Penn Treebank [11], and 450,000 sentences (with automatically induced sentence boundaries) taken from the North American News Corpus (NANC) [6]. This introduced three additional hyper-parameters specifying integer mixing ratios between the collection, WSJ, and NANC corpora.

The importance of sentence boundaries is that bigram statistics were not collected across them, which also differs from previous work. Phrase-based statistics can be expected to perform best when not collected or applied across sentential boundaries, especially as phrase length increases. This issue has largely been ignored in previous work since sentences tend to be rather long in text (maybe around 30 words in a typical newspaper), and so error introduced for short phrase statistics by approximating the entire document as a single sentence is somewhat limited. Our attention to sentence boundaries stems primarily from the fact that we are eventually interested in comparing the efficacy of bigrams with syntactic bi-lexical dependencies induced from sentences, effectively revisiting previous work [4, 10, 14] with a more accurate parser [12]. It is also worth noting that the manual summary sentences were quite short on average, suggesting their boundaries are more important to recognize in order to collect accurate phrasal statistics. Manual summaries were automatically segmented into sentences using Ratnaparkhi's tool [18]. Keywords, which could be phrasal, were already delimited in the distributed collection, and we treated each keyword phrase as its own sentence. Noting multiple spaces in the ASR transcripts appeared to correlate with sentential boundaries, we inferred these spaces corresponded to recognizer segments and used them as sentence boundaries (there is no evidence for our inference in the track's documentation). The resulting "sentences" are much longer than typical sentence-like units (SUs) found in conversational speech [8] and probably have reasonable precision but poor recall, though it is not possible to measure this without reference transcripts. As such, we used them as an expedient and left application of more accurate SU-boundary detection for future work [7].

We also applied pseudo-relevance feedback and found it significantly improved our results. Experimental parameters included the number of documents to use for feedback, and a multiplicative scaler for the original query counts. We never performed more than one iteration of feedback, nor did we try restricting term harvesting to a subset of terms rather than the entire document. This feedback scheme was developed in a short amount of time and leaves much room for improvement.

As indicated in the introduction, time constraints caused us to give far less attention to Czech than to English. We did not model bigrams or perform collection smoothing, and we used as the off-the-shelf Indri system as our model (with one minor change to its parsing code to not split tokens on accented characters). Indri actually implements a multi-Bernoulli model rather than a true unigram model, but results are roughly comparable [13]. Anecdotally comparing our unigram to Indri for the English manual data on development set topics (§4), we saw a similar trend previously reported in which the unigram model scored about a half point higher MAP [13]. In terms of stemming Czech, we evaluated use of off-the-shelf "light" and "aggressive" stemmers [20]. We also evaluated use of Indri's pseudo-relevance feedback mechanism.

4 Evaluation

This section describes our evaluation on English and Czech, including evaluation framework, parameter settings, and results.

4.1 English

The same 63 training topics and 33 evaluation topics used in 2006 were used again in 2007. Relevance assessments for the test set were distributed to participants following the 2006 evaluation, so some participants may have run error analysis experiments on the test data then without realizing it would be reused this year. Since relevant assessments for all queries were available, 2007 track organizers clearly indicated which queries comprised the test set to help ensure no one accidentally tuned on it this year. In order to preserve the test set for future use, we have performed no error analysis of our results on it. While working on the development set, we found a small problem with the distributed relevance assessments (across all topics) that some documents marked as relevant were not actually in the distributed collection, but these cases were automatically filtered out without compromising the assessments.

For the "manual" case, we used the manual summaries and keywords; we made no use pre-interview questionnaire information, or the VHF ontology. For "automatic" retrieval, we used the ASR2006B transcripts and both sets of automatic keywords. As with the manual case, we found use of the automatic keywords to improve retrieval, although we did not evaluate use of one set of keywords versus the other. We also did not evaluate other versions of the automatic transcripts, and this may be worth revisiting since some teams have reported better retrieval accuracy with the 2004 transcripts. Following previous work [26], μ was fixed at 2000 for all runs, manual and automatic. In hindsight, it would have been interesting to explore alternative μ settings since the manual and automatic "documents" used here are rather different from one another, and both differ significantly from previous retrieval experiments which varied μ on text collections. For both manual and automatic runs, best performance was almost always seen with μ_1 set to 1, and so this parameter was also largely fixed. Additional detail on parameter settings used and corresponding results on the development set can be found in the appendix (§7).

Results in Table 1 show performance of our five submitted runs on development and test sets; queries used were: title-only (T), title and description (TD), and title, description, and narrative (TDN). Representative strong results achieved the CL-SR tracks are also shown, though it should be noted that our results on the development set correspond to optimal tuning on those queries whereas the CL-SR'05 numbers do not. Retrieval accuracy was measured using mean-average

Collection	Queries	Dev MAP	CL-SR'05	Test MAP	CL-SR'06	CL-SR'07
Manual	TDN	.3829	-	.2870	.2902	?
	TD	.3443	.3129	.2366	.2710	?
	T	.3161	-	.2348	.2489	?
Auto	TDN	.1623	.2176	.0910	.0768	?
	TD	.1397	.1653	.0785	.0754	.0855

Table 1: Mean-average precision (MAP) retrieval accuracy of English submitted runs. CL-SR columns indicate representative strongs result achieved in that year's track on the same query set [15]. Statistical significance analysis between participant submissions is not possible here due to unavailability of participants' document rankings; refer to the CL-SR'07 track report for such analysis.

precision (MAP) as reported by the `trec_eval` tool version 8.1[2].

After submitting our official runs, we discovered a system bug which caused some query topic fields to be prematurely truncated. The bug did not affect the development set but did affect the *narrative* field of three test queries. After fixing the bug, we re-ran our two submissions affected by it (one manual, one automatic). We then made our only use of the test set relevance assessments to evaluate the resulting retrieval accuracy with the bug fix. The difference was substantial, and results given in the body of Table 1 show system performance with the bug fix. Without the fix, `Manual-TDN` on the test set was .2577 and `Auto-TDN` was .0831.

Regarding the impact of our enhancements to the basic unigram (bigram mixing and collections smoothing), we refer the reader to results on the development set shown in the appendix (§7). To broadly summarize, while the best unigram result was often not too far below the best bigram result, bigram results were more robust across parameter settings. Bigram statistics also appeared to have greater impact with pseudo-relevance feedback than without. As for collection smoothing, it clearly provided a substantial improvement. WSJ smoothing always helped, and NANC smoothing almost always helped, though less so in the case of pseudo-relevance feedback.

Overall, we are fairly pleased with our system's relative performance in the evaluation, but room is certainly left for improvement.

4.2 Czech

Czech retrieval accuracy was measured using the mGAP metric and tool, which mapped retrieved segments to replay times [15]. We used the distributed quickstart segments and made no use of the segmentation scripts. 29 topics used for evaluation in 2006 comprised our development set. The evaluation was blind, with a test set of 42 topics. Results are given in Table 2. Preliminary results for 2007 indicate our performance was relatively poor. While this is of course disappointing, it is not too surprising given our focus on developing a strong English baseline system. As for last year's scores, it is difficult to compare this year's results to them because they apparently suffered from a bug in the 2006 distributed quickstart segments. For the same topics, the absolute value of all teams' (preliminary) results this year far surpassed 2006 scores.

5 Conclusion

Brown's participation in the Cross-Language Speech Retrieval (CL-SR) track at the 2007 Cross Language Evaluation Forum (CLEF) represented our group's first effort in developing and applying an information retrieval (IR) system for human language. Our goal was to develop a strong baseline system which we could build on and compare to in future research. Our English system applied a novel form of Dirichlet bigram smoothing and showed the importance of collection smoothing

[2] http://trec.nist.gov/`trec_eval`

Stemmer	Feedback	Dev mGAP	Test mGAP
none	no	.0134	-
aggressive		.0140	-
light		.0144	-
none	yes	.0135	.0052
aggressive		.0146	.0106
light		.0161	.0114

Table 2: Mean Generalized Average Precision (mGAP) retrieval accuracy on Czech for the development and test sets. All runs were on the quickstart ASR transcripts using TD queries. Runs using feedback comprised our official submissions. No statistical significance testing was performed. Preliminary results for 2007 indicate the top-performing TD run achieved .0228 mGAP.

with a small collection. This system performed well, though we expect there is still room for improvement through more tuning of existing methodology. Our Czech system applied a simple unigram model under two forms of stemming and provided us with initial experience in non-English retrieval.

6 Acknowledgments

This work was initiated at the Institute of Formal and Applied Linguistics (ÚFAL)[3] at Charles University in Prague, where the first author was graciously hosted for the first half of 2007. Support for this work was provided by NSF PIRE Grant No OISE-0530118 and DARPA GALE contract HR0011-06-2-0001. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the supporting agencies.

References

[1] Thorsten Brants. Natural Language Processing in Information Retrieval. In *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands*, 2003.

[2] Seyit Ahmet Çamtepe, Mark K. Goldberg, Malik Magdon-Ismail, and Mukkai Krishn. Detecting conversing groups of chatters: a model, algorithms, and tests. In *Proceedings of the IADIS International Conference on Applied Computing*, pages 89–96, 2005.

[3] Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, 2004.

[4] Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177, 2004.

[5] J. Garofolo, G. Auzanne, and E. Voorhees. The trec spoken document retrieval track: A success story. In *the Ninth Text Retrieval Conference (TREC-9)*, 1999.

[6] David Graff. *North American News Text Corpus*, 1995. Linguistic Data Consortium. LDC95T21.

[3] http://ufal.mff.cuni.cz

[7] Mary Harper, Bonnie Dorr, John Hale, Brian Roark, Izhak Shafran, Matthew Lease, Yang Liu, Matthew Snover, Lisa Yung, Anna Krasnyanskaya, and Robin Stewart. *2005 Johns Hopkins Summer Workshop Final Report on Parsing and Spoken Structural Event Detection.*

[8] LDC. Simple metadata annotation specification version 6.2. Technical report, 2004.

[9] Matthew Lease. Natural language processing for information retrieval: the time is ripe (again). In *Proceedings of the 1st Ph.D. Workshop at the ACM Conference on Information and Knowledge Management (PIKM)*, 2007. To appear.

[10] Changki Lee, Gary Geunbae Lee, and Myung Gil Jang. Dependency structure applied to language modeling for information retrieval. *ETRI Journal*, 28:337–346, 2006.

[11] M. Marcus et al. Building a large annotated corpus of English: The Penn Treebank. *Comp. Linguistics*, 19(2):313–330, 1993.

[12] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, 2006.

[13] Donald Metzler, Victor Lavrenko, and W. Bruce Croft. Formal multiple-bernoulli models for language modeling. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 540–541, 2004.

[14] Ramesh Nallapati and James Allan. Capturing term dependencies using a language model based on sentence trees. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 383–390, 2002.

[15] D. Oard et al. Overview of the CLEF-2006 cross-language speech retrieval track. In *Working Notes for the Cross Language Evaluation Forum 2006 Workshop*, 2006.

[16] Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman, James Mayfield, Liliya Kharevych, and Stephanie Strassel. Building an information retrieval test collection for spontaneous conversational speech. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, 2004.

[17] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference*, pages 275–281, 1998.

[18] Jeffrey C. Reynar and Adwait Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the fifth conference on Applied natural language processing*, pages 16–19, 1997.

[19] S.E. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life: Okapi at trec. *Information Processing and Management*, 36(1):95–108, January 2000.

[20] J. Savoy and S. Abdou. Unine at clef-2006: experiments with monolingual, bilingual and domain-specific and robust retrieval. In *Proceedings of the Cross-Language Evaluation Forum (CLEF)*, 2006.

[21] Amit Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.*

[22] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference*, pages 21–29, 1996.

[23] Fei Song and W. Bruce Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management (CIKM)*, pages 316–321, 1999.

[24] K. Sparck Jones, S. Walker, and S.E. Robertson. A probabilistic model of information retrieval: development and comparative experiments (parts i and ii). *Information Processing and Management*, 36:779–840, 2000.

[25] Chengxiang Zhai. A brief review of information retrieval models. Technical report, Department of Computer Science, University of Illinois at Urbana-Champaign, 2007.

[26] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

7 Appendix

This section provides some additional detail on English retrieval experiments, including parameter settings and results on the development set.

7.1 Manual without pseudo-relevance feedback

MAP retrieval accuracy on development corpus for top 10 manual runs without pseudo-relevance feedback. μ_2 unigram weight was either 500 or 100000 (the latter effectively nullifying bigram statistics). Use of manual key words in addition to manual summaries was varied. Collection smoothing was varied between: none, WSJ, and WSJ+NANC (simple concatenation). Overall, use of manual keywords is seen to help though not dramatically. Collection smoothing, in contrast, appears to be much more important.

```
0.3369  TDN 500 with-manual-keywords.wsj+nanc
0.3312  TDN 500 wsj
0.3269  TDN 500 wsj+nanc
0.3260  TDN 500 with-manual-keywords.wsj
0.3236  TDN 100000 with-manual-keywords.wsj+nanc
0.3218  TDN 100000 wsj
0.3149  TDN 100000 with-manual-keywords.wsj
0.3116  TDN 100000 wsj+nanc
0.2967  TDN 500 no-collection-smoothing
0.2914  TDN 500 with-manual-keywords.no-collection-smoothing
...
0.3091  TD 500 with-manual-keywords.wsj+nanc
0.3058  TD 500 wsj
0.3033  TD 500 with-manual-keywords.wsj
0.3021  TD 100000 wsj
0.3016  TD 500 wsj+nanc
0.2992  TD 100000 with-manual-keywords.wsj+nanc
0.2965  TD 100000 with-manual-keywords.wsj
0.2945  TD 100000 wsj+nanc
0.2852  TD 500 with-manual-keywords.no-collection-smoothing
0.2825  TD 500 no-collection-smoothing
...
0.2721  T 500 with-manual-keywords.wsj+nanc
0.2716  T 100000 with-manual-keywords.wsj+nanc
0.2689  T 100000 wsj
0.2663  T 500 with-manual-keywords.wsj
0.2662  T 500 wsj+nanc
```

```
0.2661  T 100000 wsj+nanc
0.2656  T 500 wsj
0.2646  T 100000 with-manual-keywords.wsj
0.2611  T 500 no-collection-smoothing
0.2605  T 100000 no-collection-smoothing
```

7.2 Manual with pseudo-relevance feedback

MAP retrieval accuracy for top 10 parameter settings on development corpus for each query condition given manual summary and keywords with pseudo-relevance feedback. Fields are: query condition, μ_2 unigram weight, number of documents to use for feedback, and mixing weight of original query relative to feedback set. μ_1 bigram weight was fixed at 1. The relative rarity of μ_2 unigram weight parameter setting of 100000 suggests the bigram model provides more robust, strong performance for medium and long queries.

```
0.3829  TDN.2000.20.8
0.3829  TDN.100000.20.4
0.3823  TDN.2000.15.4
0.3821  TDN.2000.15.3
0.3821  TDN.1000.20.8
0.3817  TDN.1500.20.8
0.3812  TDN.1500.15.4
0.3812  TDN.100000.15.4
0.3811  TDN.2000.20.4
0.3811  TDN.1500.15.3
...
0.3443  TD.2000.15.8
0.3440  TD.2000.10.8
0.3437  TD.1500.10.8
0.3435  TD.100000.15.8
0.3431  TD.2000.10.4
0.3431  TD.1500.15.8
0.3430  TD.1000.10.8
0.3425  TD.750.10.8
0.3425  TD.2000.15.16
0.3425  TD.2000.10.3
...
0.3131  T.2000.20.16
0.3120  T.1500.20.16
0.3116  T.1500.15.16
0.3116  T.100000.20.16
0.3110  T.2000.15.16
0.3110  T.1000.20.16
0.3107  T.2000.20.8
0.3102  T.100000.20.8
0.3100  T.100000.15.16
0.3098  T.100000.20.4
```

7.3 Automatic without pseudo-relevance feedback

MAP retrieval accuracy for top 10 parameter settings on development corpus for each query condition given ASR06B recognizer output plus automatic keywords and no pseudo-relevance feedback. NANC-smoothing consistently appeared in the two shorter query conditions while rarely appearing for verbose queries. Fields are: query condition, μ_1 bigram weight, μ_2 unigram weight, collection weight, WSJ weight, and NANC weight.

```
0.1396  TDN.1.750.coll+akw-1.wsj-16.nanc-0
0.1395  TDN.1.750.coll+akw-1.wsj-8.nanc-0
0.1393  TDN.1.750.coll+akw-1.wsj-8.nanc-1
0.1393  TDN.1.500.coll+akw-1.wsj-8.nanc-0
0.1393  TDN.1.1250.coll+akw-1.wsj-16.nanc-0
0.1392  TDN.1.1500.coll+akw-1.wsj-16.nanc-0
0.1391  TDN.1.1000.coll+akw-1.wsj-8.nanc-0
0.1391  TDN.1.1000.coll+akw-1.wsj-16.nanc-0
0.1390  TDN.1.1250.coll+akw-1.wsj-8.nanc-0
0.1389  TDN.1.750.coll+akw-1.wsj-4.nanc-1
...
0.1240  TD.1.1500.coll+akw-2.wsj-1.nanc-1
0.1237  TD.1.1250.coll+akw-2.wsj-1.nanc-1
0.1232  TD.1.2000.coll+akw-1.wsj-1.nanc-1
0.1231  TD.1.1500.coll+akw-1.wsj-1.nanc-1
0.1231  TD.1.1250.coll+akw-1.wsj-2.nanc-1
0.1230  TD.1.2000.coll+akw-1.wsj-2.nanc-1
0.1230  TD.1.1250.coll+akw-1.wsj-1.nanc-1
0.1229  TD.1.1500.coll+akw-1.wsj-2.nanc-1
0.1229  TD.1.1000.coll+akw-1.wsj-4.nanc-1
0.1229  TD.1.1000.coll+akw-1.wsj-2.nanc-1
...
0.1103  T.1.5000.coll+akw-2.wsj-1.nanc-1
0.1103  T.1.2000.coll+akw-2.wsj-1.nanc-1
0.1102  T.1.5000.coll+akw-1.wsj-2.nanc-1
0.1102  T.1.5000.coll+akw-1.wsj-1.nanc-1
0.1099  T.1.5000.coll+akw-1.wsj-4.nanc-1
0.1099  T.1.2000.coll+akw-1.wsj-4.nanc-0
0.1099  T.1.1500.coll+akw-2.wsj-1.nanc-1
0.1099  T.1.10000.coll+akw-1.wsj-1.nanc-1
0.1097  T.5.5000.coll+akw-2.wsj-1.nanc-1
0.1097  T.1.2000.coll+akw-1.wsj-1.nanc-1
```

7.4 Automatic with pseudo-relevance feedback

MAP retrieval accuracy for top 10 parameter settings on development corpus for each query condition given ASR06B recognizer output plus automatic keywords using pseudo-relevance feedback. WSJ-smoothing was consistently useful while NANC smoothing was not. Fields are: query condition, μ_2 unigram weight, number of documents to use for feedback, mixing weight of original query relative to feedback set, collection weight, WSJ weight, and NANC weight. μ_1 bigram weight was fixed at 1.

```
0.1623  TDN.1250.10.20.coll-akw-1.wsj-4.nanc-0
0.1617  TDN.750.5.20.coll-akw-1.wsj-2.nanc-0
0.1616  TDN.1500.10.20.coll-akw-1.wsj-8.nanc-0
0.1616  TDN.1500.10.20.coll-akw-1.wsj-2.nanc-0
0.1615  TDN.750.5.20.coll-akw-1.wsj-4.nanc-0
0.1615  TDN.2000.10.20.coll-akw-1.wsj-4.nanc-0
0.1615  TDN.1250.10.20.coll-akw-1.wsj-8.nanc-0
0.1613  TDN.750.10.20.coll-akw-1.wsj-4.nanc-0
0.1613  TDN.750.10.20.coll-akw-1.wsj-1.nanc-0
0.1611  TDN.2000.10.20.coll-akw-1.wsj-1.nanc-1
...
0.1397  TD.2000.10.20.coll-akw-2.wsj-1.nanc-1
```

```
0.1396  TD.2000.10.20.coll-akw-1.wsj-1.nanc-1
0.1394  TD.2000.10.20.coll-akw-1.wsj-8.nanc-0
0.1394  TD.1250.10.20.coll-akw-2.wsj-1.nanc-1
0.1394  TD.100000.15.20.coll-akw-1.wsj-16.nanc-1
0.1393  TD.2000.10.20.coll-akw-1.wsj-2.nanc-1
0.1393  TD.1500.10.20.coll-akw-2.wsj-1.nanc-1
0.1393  TD.100000.15.20.coll-akw-1.wsj-16.nanc-0
0.1390  TD.1500.10.20.coll-akw-1.wsj-4.nanc-0
0.1389  TD.2000.10.20.coll-akw-4.wsj-1.nanc-1
...
0.1242  T.100000.10.20.coll-akw-1.wsj-16.nanc-0
0.1240  T.2000.10.20.coll-akw-1.wsj-2.nanc-0
0.1240  T.2000.10.20.coll-akw-1.wsj-1.nanc-0
0.1237  T.100000.10.20.coll-akw-1.wsj-8.nanc-0
0.1236  T.2000.10.15.coll-akw-1.wsj-1.nanc-0
0.1235  T.1500.10.20.coll-akw-4.wsj-1.nanc-0
0.1234  T.2000.10.15.coll-akw-2.wsj-1.nanc-0
0.1233  T.2000.10.20.coll-akw-4.wsj-1.nanc-1
0.1233  T.1500.10.20.coll-akw-1.wsj-1.nanc-0
0.1233  T.100000.10.20.coll-akw-1.wsj-8.nanc-1
```