A Dirichlet-smoothed Bigram Model for Retrieving Spontaneous Speech

Matthew Lease and Eugene Charniak

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

Abstract. We present two simple but effective smoothing techniqes for the standard language model (LM) approach to information retrieval [12]. First, we extend the unigram Dirichlet smoothing technique popular in IR [17] to bigram modeling [16]. Second, we propose a method of *collection expansion* for more robust estimation of the LM prior, particularly intended for sparse collections. Retrieval experiments on the MALACH archive [9] of automatically transcribed and manually summarized spontaneous speech interviews demonstrates strong overall system performance and the relative contribution of our extensions[1].

1 Introduction

In the language model (LM) paradigm for information retrieval (IR), a document's relevance is estimated as the probability of observing the query string as a random sample from the document's underlying LM [12]. The standard unigram LM approach has been shown to have a strong theoretical connection to TF-IDF [17] and comparable performance to other state-of-the-art approaches like vector similarity and the "probabilistic" approach [1]. This paper presents two modest smoothing-based extensions in the LM paradigm.

Whereas the unigram model and other standard approaches to retrieval typically assume bag-of-words independence between terms, modeling even a simple notion of term dependency represents a useful step toward richer modeling of queries and documents. Previous work in bigram modeling provided a valuable first step in this direction within the LM paradigm and demonstrated its empirical merit [16]. Subsequent to this, Dirichlet smoothing with unigram models was found to elegantly and effectively capture the intuition that longer documents should require less smoothing since they provide more support for the maximum-likelihood (ML) estimate [17]. While one would expect bigram models could similarly benefit, we have not seen a Dirichlet-smoothed bigram model described or evaluated in the IR literature. Consequently, we describe such a model here and report on its effectiveness. As with the earlier bigram formulation [16], our approach easily generalizes to higher-order mixtures.

[1] An earlier version of this work was presented in the CLEF 2007 Working Notes.

The second extension we describe addresses smoothing at the collection-level. As suggested above, smoothing plays an important role in inferring accurate document LMs, and it can be accomplished in a principled manner via *maximum a posteriori* (MAP) estimation using a prior model. For IR, the prior is typically estimated from collection statistics, but just as estimating a robust document model is often challenging due to document sparsity, estimating the prior from a small (i.e. sparse) collection can be equally problematic. To address this, we propose estimating the prior from an "expanded" version of the collection containing additional statistics drawn from external corpora. This idea closely parallels previous work expanding documents with similar ones found in external sources [15]. Here, collection-wide statistics are expanded via external corpora to enable more robust estimation of the LM prior. We show simple collection expansion via broad English corpora significantly improves retrieval accuracy.

We evaluated our model and extensions via retrieval experiments on the MALACH archive of automatically transcribed and manually summarized spontaneous speech interviews [9]. These experiments were conducted as part of the Cross-Language Speech Retrieval track's shared task [11] at the 2007 Cross Language Evaluation Forum. Results show the overall competitive performance of our system as well as the relative contribution of our extensions.

The remainder of our paper is presented as follows: methodology is discussed in §2, relevant details of the MALACH collection and pre-processing are described in §3, evaluation procedure and results are presented in §4, and §5 summarizes and describes future work.

2 Method

2.1 Dirichlet-smoothed Bigram Modeling

The link recently forged between language modeling and information retrieval [12] established a new mathematical foundation for IR that made a large body of existing theoretical knowledge and empirical experience suddenly applicable. This connection opened the door to an exciting new line of IR research that has already delivered new theoretical insights and excellent empirical results, while at the same time leaving open many interesting directions to pursue.

The core insight of the LM approach is that rather than trying to directly connect a query to its relevant documents by measuring similarity of observed terms, we instead seek an indirect connection by inferring a common underlying stochastic distribution from which query and document arise. The key challenges in this approach are hypothesizing the form of the underlying source models and finding an effective estimation procedure given the brevity of observed evidence.

If we assume *a priori* that all documents are equally likely to be relevant to a given query, then by Bayes inversion we can formulate the document ranking task as estimating query Q's likelihood under each document D's underlying LM: $P(D|Q) \propto P(Q|D)$. Further assuming complete independence between observed terms (naive Bayes) yields a bag-of-words unigram model in which query

likelihood is estimated by the product of individual term probabilities under the document LM $P(\cdot|D)$.

How do we estimate this model $P(\cdot|D)$? One option is ML. Assuming vocabulary size V, word w_i occurring in D with frequency f_{w_i}, and $P(\cdot|D)$ being parameterized by Θ, we could seek the particular $\widehat{\Theta}$ maximizing D's likelihood

$$P(D|\Theta) = \prod_{i=1}^{V} {\theta_i}^{f_{w_i}} \tag{1}$$

which would be the assignment to Θ respecting the empirical frequencies f. However, such use of ML is problematic in that a single unobserved query term would completely nullify query likelihood, making the entire framework exceedingly fragile. The problem here is that in observing only a small sample (i.e. a brief document) from an underlying distribution, effects of chance variation will be prominent and distort sample statistics away from those governing the generating distribution. Fortunately, prior knowledge about the distribution can be leveraged in a principled way via MAP estimation. *A priori*, we might reasonably assume $P(\cdot|D)$ should resemble the collection's *average* document model $P(\cdot|C)$. This, in turn, could be estimated via ML by summing statistics across all documents, which generally do provide sufficient evidence for a robust estimate.

Such prior knowledge can be elegantly incorporated into a language model via the Dirichlet distribution, specified by hyperparameters $\alpha > 0$ and defining a distribution over multinomial parameterizations $P(\Theta;\alpha)$ [6]. For the unigram model defined above, the corresponding Dirichlet prior would be defined as

$$P(\Theta;\alpha) \doteq Dir(\alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^{V} {\theta_i}^{\alpha_i - 1} \tag{2}$$

where $Z(\alpha)$ denotes normalization. This prior is particularly convenient for MAP estimation because its distribution is conjugate to the multinomial, meaning the posterior will also be Dirichlet. Hence, combining likelihood (1) and prior (2):

$$P(\Theta|D;\alpha) \propto P(\Theta;\alpha)P(D|\Theta) \propto \prod_{i=1}^{V} {\theta_i}^{\alpha_i - 1} \prod_{i=1}^{V} \theta_i^{f_{w_i}} = \prod_{i=1}^{V} \theta_i^{f_{w_i} + \alpha_i - 1} \tag{3}$$

A true Bayesian would next compute the predictive distribution over Θ, but we will instead assume a peaked posterior and find the single most-likely $\widehat{\Theta}$ to explain our data via the maximum approximation. Comparing our likelihood and posterior equations (1) and (3), we can see that maximizing the posterior is quite similar to maximizing the likelihood, only the data now consists of both the empirical evidence and "pseudo" α observations. In other words, the posterior maximum is simply the combined relative frequency of the observed and pseudo data. Finally, letting $\alpha - 1 = \mu P(\cdot|C)$ for $\mu >= 0$, we see our empirical document statistics are smoothed with μ pseudo-counts drawn from our average document model $P(\cdot|C)$ to yield IR's popular Dirichlet-smoothed unigram model [17]

$$P(w|D,C) = \frac{f_w + \mu P(w|C)}{N + \mu} \tag{4}$$

where N specifies the length of D. The attractiveness of this smoothing strategy lies in the fact that as document length increases, providing more evidence for the ML estimate, the impact of the prior model will correspondingly diminish.

To extend this strategy to bigram modeling, we similarly smooth the empirical bigram estimate with hyperparameter μ_1 pseudo-counts distributed fractionally according to the collection prior bigram model, $P(w_i|w_{i-1}, C)$:

$$P(w_i|w_{i-1}, D, C) = \frac{f_{w_{i-1},w_i} + \mu_1 P(w_i|w_{i-1}, C)}{f_{w_{i-1}} + \mu_1} \tag{5}$$

Unigram and bigram models can then be easily mixed by treating our smoothed unigram distribution $P(w|D, C)$ as an additional prior on the bigram model and adding in μ_2 pseudo-counts drawn from it:

$$P(w_i|w_{i-1}, D, C) = \frac{f_{w_{i-1},w_i} + \mu_1 P(w_i|w_{i-1}, C) + \mu_2 P(w|D, C)}{f_{w_{i-1}} + \mu_1 + \mu_2} \tag{6}$$

Whereas earlier work inferred the hyperparameters α from data in order to realize a coupled prior tying unigram and bigram models [6], our formulation can be viewed as a less sophisticated alternative that reduces α to three hyperparameters, μ, μ_1, and μ_2, to be tuned on development data.

2.2 Collection Expansion

The second extension we describe addresses more robust estimation of the LM prior by performing smoothing at the collection-level. As discussed above, ML estimation of document LMs is hurt by document sparsity, and hence MAP estimation is commonly employed instead using an informative prior induced from the collection. The effectiveness of this strategy, however, relies on accurate estimation of the prior, which can be challenging for small (i.e. sparse) collections.

To address this, we propose estimating the prior from an "expanded" version of the collection containing additional data drawn from external corpora. This approach parallels traditional work in document expansion in which collection documents are expanded with external, related documents [15]. In both cases, the underying idea of expansion being employed is characteristic of a broad finding in the learning community that having additional similar data enables more robust estimation. In our case of *collection expansion*, we hope to compensate for collection sparsity by drawing upon "similar" data from external corpora.

For this work, we simply leveraged two broad English newspaper corpora: the Wall Street Journal (WSJ) and the North American News Corpus (NANC) [2]. Specifically, we expanded the collection as a linear mixture with 40K sentences (830K words) from WSJ (as found in the Penn Treebank [7]) and 450K sentences (9.5M words) from NANC, with tunable hyperparameters specifying integer mixing ratios between corpora. The particular corpora and mixing scheme used could likely be improved by a more sophisticated strategy. For example, results in §4 show significant improvement for modeling manually-written summaries but not

for automatic transcriptions, likely due to mismatch between the external corpora and the automatic transcriptions. Bigram statistics in expansion corpora were not collected across sentence boundaries, which were manually annotated in WSJ and automatically detected in NANC [8].

3 Data

This section describes the retrieval collection used and pre-processing performed. A more complete description of the collection can be found elsewhere [9–11].

Data used came from the Survivors of the Shoah Visual History Foundation (VHF) archive of interviews with Holocaust survivors, rescuers, and witnesses. A subset of this archive was manually and automatically processed by VHF and members of the MALACH initiative (Multilingual Access to Large Spoken Archives) in order to improve access to this archive and other such collections of spontaneous speech content. As part of this effort, interviews were manually segmented and summarized, as well as automatically transcribed (several variant transcriptions were produced). Manual transcription was limited and not provided for interviews included in the retrieval collection. Each interview segment was also manually assigned a set of keywords according to a careful ontology developed by VHF, and two versions of of automatically detected keywords were also provided. Topics used for retrieval were based on actual information requests received by VHF from interested parties and were expressed in typical TREC-style with increasingly detailed title, description, and narrative fields [9].

In terms of pre-processing, sentence boundaries were automatically detected to collect more accurate bigram statistics. Boundaries for manual summaries were detected using a standard tool [13] and interview segment keyword phrases were each treated as separate sentences. We noted the presence of multiple contiguous spaces in automatic transcriptions appeared to correlate with sentence-like units (SUs) [3] and so segmented sentences based on them[2]. Use of automatic SU-boundary detection is left for future work [14].

4 Evaluation

This section describes system evaluation, including experimental framework, parameter settings, and results. Retrieval experiments were performed as part of the 2007 Cross Language Evaluation Forum's Cross-Language Speech Retrieval (CL-SR) task [11].

We used 25 topics for development and 33 for final testing (the 2005 and 2006 CL-SR evaluation sets, respectively; the 2006 test set was re-used for the 2007 evaluation). For the "manual" retrieval condition, segments consisted of manual summaries and keywords. For the "automatic" condition, we used the ASR2006B transcripts and both versions of automatic keywords. Following previous work [17], the unigram Dirichlet smoothing parameter μ was fixed at 2000

[2] Collection documentation does not discuss this.

for both manual and automatic conditions. Best performance was usually observed with μ_1 set to 1, while optimal μ_2 settings varied.

A limited pseudo-relevance feedback (PRF) scheme was also employed. As in standard practice, documents were ranked by the model according to the original query, with the most likely documents taken to comprise its feedback set (the number of feedback documents used varied). The query was then reformulated by adding the 50 most frequent bigrams from each feedback document. A tuning parameter specified a multiplier for the original query counts to provide a means of weighting the original query relative to the feedback set. This scheme likely could be improved by separate treatment for unigram feedback and weighting feedback documents by document likelihood under the original query.

Results in Table 1 show performance of our five official runs on development and test sets[3]; queries used were: title-only (T), title and description (TD), and title, description, and narrative (TDN). Representative strong results achieved in 2007's and previous years' CL-SR tracks [10, 11] are also shown, though it should be noted that our results on the development set correspond to tuning on those queries whereas the CL-SR'05 official results do not. Retrieval accuracy was measured using mean-average precision reported by `trec_eval` version 8.1[4].

Collection	Queries	Dev	CL-SR'05	Test	CL-SR'06	CL-SR'07
Manual	TDN	.3829	-	.2870	.2902	.2847
	TD	.3443	.3129	.2366+	.2710	.2761+
	T	.3161	-	.2348	.2489	-
Auto	TDN	.1623	.2176	.0910	.0768	-
	TD	.1397	.1653	.0785-	.0754	.0855-

Table 1. Mean-average precision retrieval accuracy of submitted runs. CL-SR columns indicate representative strong results achieved in that year's track on the same query set [10, 11]. Runs marked above with `+/-` were reported in the 2007 track report to represent statistical significance and non-significance, respectively.

Table 2 shows the impact of our extensions compared to the baseline Dirichlet-smoothed unigram retrieval model for the no-PRF "manual" condition. Of the two extensions, collection expansion is seen to have greater effect, with the combination yielding the best result. The effect of the extensions with the "automatic" condition was marginal (the best absolute improvement seen was 0.3% achieved by the bigram model). With collection expansion, we suspect this is due to the mismatch between the collection's spontaneous speech and the text corpora used for expansion (§2), and we plan to investigate use of better matched corpora in

[3] Following submission of official runs, we found a bug affecting our parsing of the *narrative* field of three test queries. Table 1 show system performance with the bug fixed. Without the fix, `Manual-TDN` on the test set was .2577 and `Auto-TDN` was .0831.

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

Model	T	TD	TDN
Unigram baseline	.2605	.2722	.2810
Dirichlet bigram	.2545 (-2.3%)	.2852 (4.8%)	.2967 (5.6%)
Collection Expansion	.2716 (4.3%)	.3021 (11.0%)	.3236 (15.2%)
Combination	.2721 (4.5%)	.3091 (13.6%)	.3369 (19.9%)

Table 2. Improvement in mean-average precision on the development set over the unigram baseline model for Dirichlet-smoothed bigram modeling and collection expansions, alone and in combination (manual condition, no pseudo-relevance feedback).

future work. As for the bigram model, automatic transcription noise is more problematic than with unigrams since recognition error further impacts prediction of subsequent terms. One strategy for addressing this would be to work off the recognition lattice instead of the one-best transcription. Another challenge to the bigram model is the presence of disfluency in spontaneous speech, which disrupts bigram statistics. Automatic detection and deletion of disfluency could help address this and thereby also render the spoken document more amenable to smoothing via external text corpora [5].

For manual retrieval with PRF, the combination of extensions was used in selecting the set of documents for feedback. For PRF runs using this feedback set, the extensions were seen to provide minimal further benefit, with PRF tuning parameters dominating the variance in performance observed. Since PRF produces a query more tailored to collection statistics, expanded collection statistics may be less useful in PRF settings.

5 Conclusion and Future Work

This paper presented two smoothing-based extensions to the standard language model approach to information retrieval: Dirichlet-smoothed bigram modeling and collection expansion. Empirical results demonstrated the relative contribution of the extensions and competitive overall system performance.

Future work will explore two lines of research in LM-based information retrieval [4]: inferring latent structure to derive richer representations for modeling, and revisiting existing SDR retrieval methodology with greater attention to modeling spontaneous speech phenomena.

Acknowledgments

The authors thank Will Headden and our anonymous reviewers for their valuable comments. This work was initiated by the first author while hosted at the Institute of Formal and Applied Linguistics (ÚFAL) at Charles University in Prague. 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. Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. In *SIGIR '04: Proc. 27th ACM SIGIR conference*, pages 49–56, 2004.
2. David Graff. *North American News Text Corpus*, 1995. Linguistic Data Consortium. LDC95T21.
3. LDC. Simple metadata annotation specification 6.2. Technical report, 2004.
4. 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.
5. Matthew Lease, Mark Johnson, and Eugene Charniak. Recognizing disfluencies in conversational speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1566–1573, September 2006.
6. D. J. C. MacKay and L. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19, 1995.
7. Michell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
8. David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the HLT-NAACL conference*, pages 152–159, 2006.
9. 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: Proc. of the 27th annual international ACM SIGIR conference*, pages 41–48, 2004.
10. Douglas W. Oard, Jianqiang Wang, Gareth Jones, Ryen White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran. Overview of the clef-2006 cross-language speech retrieval track. In *Evaluation of Multilingual and Multi-modal Information Retrieval - Seventh Workshop of the Cross-Language Evaluation Forum*, volume LNCS 4730, pages 744–758. Springer, 2007.
11. Pavel Pecina, Petra Hoffmannova, Gareth J.F. Jones, Ying Zhang, and Douglas W. Oard. Overview of the CLEF-2007 cross language speech retrieval track. In *Evaluation of Multilingual and Multi-modal Information Retrieval - Eighth Workshop of the Cross-Language Evaluation Forum*. Springer, 2008.
12. 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.
13. 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.
14. B. Roark, Yang Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung. Reranking for sentence boundary detection in conversational speech. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 545–548, 2006.
15. Amit Singhal and Fernando Pereira. Document expansion for speech retrieval. In *Proc. of the 22nd annual international ACM SIGIR conference*, pages 34–41, 1999.
16. 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.
17. 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.