Cache-Augmented Latent Topic Language Models for Speech Retrieval

Jonathan Wintrode
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD
jcwintr@cs.jhu.edu

Abstract

We aim to improve speech retrieval performance by augmenting traditional N-gram language models with different types of topic context. We present a latent topic model framework that treats documents as arising from an underlying topic sequence combined with a cache-based repetition model. We analyze our proposed model *both* for its ability to capture word repetition via the cache and for its suitability as a language model for speech recognition and retrieval. We show this model, augmented with the cache, captures intuitive repetition behavior across languages and exhibits lower perplexity than regular LDA on held out data in multiple languages. Lastly, we show that our joint model improves speech retrieval performance beyond N-grams or latent topics alone, when applied to a term detection task in all languages considered.

1 Introduction

The availability of spoken digital media continues to expand at an astounding pace. According to YouTube's publicly released statistics, between August 2013 and February 2015 content upload rates have tripled from 100 to 300 hours of video per minute (YouTube, 2015). Yet the *information* content therein, while accessible via links, tags, or other user-supplied metadata, is largely inaccessible via content search within the speech.

Speech retrieval systems typically rely on Large Vocabulary Continuous Speech Recognition (LVSCR) to generate a lattice of word hypotheses for each document, indexed for fast search (Miller and others, 2007). However, for sites like YouTube, localized in over 60 languages (YouTube, 2015), the likelihood of high accuracy speech recognition in most languages is quite low.

Our proposed solution is to focus on *topic information* in spoken language as a means of dealing with errorful speech recognition output in many languages. It has been repeatedly shown that a task like topic classification is robust to high (40-60%) word error rate systems (Peskin, 1996; Wintrode, 2014b). We would leverage the topic signal's strength for retrieval in a high volume, multilingual digital media processing environment.

The English word *topic*, defined as a particular 'subject of discourse' (Houghton-Mifflin, 1997), arises from the Greek root, $\tau o \pi o \varsigma$, meaning a physical 'place' or 'location'. However, the semantic concepts of a particular subject are not disjoint from the physical location of the words themselves.

The goal of this particular work is to jointly model two aspects of topic information, *local context* (repetition) and *broad context* (subject matter), which we previously treated in an ad hoc manner (Wintrode and Sanjeev, 2014) in a latent topic framework. We show that in doing so we can achieve better word retrieval performance than language models with only N-gram context on a diverse set of spoken languages.

2 Related Work

The use of both repetition and broad topic context have been exploited in a variety of ways by the speech recognition and retrieval communities. Cache-based or adaptive language models were

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 1–8,
Denver, Colorado, June 1, 2015.

some of the first approaches to incorporate information beyond a short N-gram history (where N is typically 3-4 words).

Cache-based models assume the probability of a word in a document d is influenced both by the global frequency of that word and N-gram context as well as by the N-gram frequencies of d (or preceding *cache* of K words). Although most words are rare at the corpus level, when they do occur, they occur in bursts. Thus a local estimate, from the *cache*, may be more reliable than the global estimate. Jelinek (1991) and Kuhn (1990) both successfully applied these types of models for speech recognition, and Rosenfeld (1994), using what he referred to as 'trigger pairs', also realized significant gains in WER. More recently, recurrent neural network language models (RNNLMs) have been introduced to capture more of these "long-term dependencies" (Mikolov et al., 2010). In terms of speech retrieval, recent efforts have looked at exploiting repeated keywords at search time, without directly modifying the recognizer (Chiu and Rudnicky, 2013; Wintrode, 2014a).

Work within the information retrieval (IR) community connects topicality with retrieval. Hearst and Plaunt (1993) reported that the "subtopic structuring" of documents can improve full-document retrieval. Topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2001) are used to the augment the document-specific language model in probabilistic, language-model based IR (Wei and Croft, 2006; Chen, 2009; Liu and Croft, 2004; Chemudugunta et al., 2007). In all these cases, topic information was helpful in boosting retrieval performance above baseline vector space or N-gram models.

Our proposed model closely resembles that from Chemudugunta et al. (2007), with our notions of broad and local context corresponding to their "general and specific" aspects. The unigram cache case of our model should correspond to their "special words" model, however we do not constrain our cache component to only unigrams.

With respect to speech recognition, Florian and Yarowsky (Florian and Yarowsky, 1999) and Khudanpur and Wu (Khudanpur and Wu, 1999) use vector-space clustering techniques to approximate the topic content of documents and augment a baseline N-gram model with topic-specific N-gram counts. Clarkson and Robinson (1997) proposed a similar application of cache and mixture models, but only demonstrate small perplexity improvements. Similar approaches use latent topic models to infer a topic mixture of the test document (soft clustering) with significant recognition error reductions (Heidel et al., 2007; Hsu and Glass, 2006; Liu and Liu, 2008; Huang and Renals, 2008). Instead of interpolating with a traditional backoff model, Chien and Chueh (2011) use topic models with and without a dynamic cache to good effect as a class-based language model.

We build on the cluster-oriented results, particularly Khudanpur and Wu (1997) and Wintrode and Khudanpur (2014), but within an explicit framework, jointly capturing both types of topic information that many have leveraged individually.

Algorithm 1 Cache-augmented generative process

for all $t \in \mathcal{T}$ **do**
  draw $\phi^{(t)} \sim \text{Dirichlet}(\beta)$
for all $d \in \mathcal{D}$ **do**
  draw $\theta^{(d)} \sim \text{Dirichlet}(\alpha)$
  draw $\kappa^{(d)} \sim \text{Beta}(\nu_0, \nu_1)$
  for $w_{d,i}, 1 \leq i \leq |d|$ **do**
    draw $k_{d,i} \sim \text{Bernoulli}(\kappa^{(d)})$
    if $k_{d,i} = 0$ **then**
      draw $z_{d,i} \sim \theta^{(d)}$
      draw $w_{d,i} \sim \phi^{(t=z_{d,i})}$
    else
      draw $w_{d,i} \sim \text{Cache}(d, W_{-i})$
    end if

3 Cache-augmented Topic Model

We propose a straightforward extension of the LDA topic model (Blei et al., 2003; Steyvers and Griffiths, 2007), allowing words to be generated *either* from a latent topic or from a document-level cache. At each word position we flip a biased coin. Based on the outcome we either generate a latent topic and then the observed word, or we pick a new word directly from the cache of already observed words. Thus we would jointly learn the underlying topics and the tendency towards repetition.

As with LDA, we assume each corpus is drawn from $\mathcal{T}$ latent topics. Each topic is denoted $\phi^{(t)}$, a

multinomial random variable in the size of the vocabulary where $\phi_v^{(t)}$ is the probability $P(w_v|t)$. For each document we draw $\theta^{(d)}$, where $\theta_t^{(d)}$ is the probability $P(t|d)$.

We introduce two additional sets of variables, $\kappa^{(d)}$ and $k_{d,i}$. The state $k_{d,i}$ is a Bernoulli variable indicating whether a word $w_{d,i}$ is drawn from the cache or from the latent topic state. $\kappa^{(d)}$ is the document specific prior on the cache state $k_{d,i}$.

Algorithm 1 gives the generative process explicitly. We choose a Beta prior $\kappa^{(d)}$ for the Bernoulli variables $k_{d,i}$. As with the Dirichlet priors, this allows for a straightforward formulation of the joint probability $P(W, Z, K, \Phi, \Theta, \kappa)$, from which we derive densities for Gibbs sampling. A plate diagram is provided in Figure 1, illustrating the dependence both on latent variables and the cache of previous observations.

We implement our model as a collapsed Gibbs sampler extending Java classes from the Mallet topic modeling toolkit (McCallum, 2002). We use the Gibbs sampler for parameter estimation (training data) and inference (held-out data). We also leverage Mallet's hyperparameter re-estimation (Wallach et al., 2009), which we apply to α, β, and ν.

4 Language Modeling

Our primary goal in constructing this model is to apply it to language models for speech recognition and retrieval. Given an LVCSR system with a standard N-gram language model (LM), we now describe how we incorporate the inferred topic and cache model parameters of a new document into the base LM for subsequent recognition tasks *on that specific document*.

We begin by estimating model parameters on a training corpus: topics $\phi^{(t)}$, cache proportions $\kappa^{(d)}$, and hyperparameters, α, β, and ν (the Beta hyperparameter). In our experiments we restrict the training set to the LVCSR acoustic and language model training. This restriction is required by the Babel task, not the model. Using other corpora or text resources certainly should be considered for other tasks.

To apply the model during KWS, we first decode a new audio document d with the base LM, P_L and extract the most likely observed word sequence W for inference. The inference process gives us the estimates for $\theta^{(d)}$ and $\kappa^{(d)}$, which we then use to compute *document-specific* and cache-augmented language models.

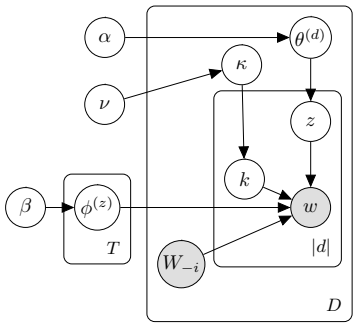

Figure 1: Cache-augmented model plate diagram.

From a language modeling perspective we treat the multinomials $\phi^{(t)}$ as unigram LM's and use the inferred topic proportions $\theta^{(d)}$ as a set of mixture weights. From these we compute the document-specific unigram model for d (Eqn. 1). This serves to capture what we have referred to as the *broad* topic context.

We incorporate both P_d as well as the cache P_c (local context) into the base model P_L using linear interpolation of probabilities. Word histories are denoted h_i for brevity. For our experiments we first combine P_d with the N-gram model (Eqn. 2). We then interpolate with the cache model to get a joint topic and cache language model (Eqn. 4).

$$P_d(w_i) = \sum_{t=1}^{T} \theta_t^{(d)} \cdot \phi_i^{(t)} \quad (1)$$

$$P_{Ld}(w_i) = \lambda P_d(w_i) + (1-\lambda) \cdot P_L(w_i) \quad (2)$$

$$P_{dc}(w_i) = \kappa^{(d)} P_c(w_i) + (1-\kappa^{(d)}) \cdot P_d(w_i) \quad (3)$$

$$P_{Ldc}(w_i|h_i) = \kappa^{(d)} P_c(w_i|h_i) + (1-\kappa^{(d)}) \cdot P_{Ld}(w_i|h_i) \quad (4)$$

We expect the inferred document cache probability $\kappa^{(d)}$ to serve as a natural interpolation weight when combining document-specific unigram model P_{dc} and cache. We consider alternatives to per-document $\kappa^{(d)}$ as part of the speech retrieval evaluation (Section 6) and can show that our model's estimate is indeed effective.

Language	50t	100t	150t	200t
Tagalog	0.41	0.29	0.22	0.16
Vietnamese	0.51	0.39	0.29	0.22
Zulu	0.33	0.26	0.21	0.16
Tamil	0.36	027	0.18	0.14

Table 1: Mean $\kappa^{(d)}$ inferred from 10 hour development data, by number of latent topics

5 Model Analysis

Before looking at the model in terms of retrieval performance (Section 6), here we aim to examine how our model captures the repetition of each corpus and how well it functions as a language model (cf. Equation 3) in terms of perplexity.

To focus on language models for speech retrieval in the limited resource setting, we build and evaluate our model under the IARPA Babel Limited Language Pack (LP), No Target Audio Reuse (NTAR) condition (Harper, 2011). We selected the Tagalog, Vietnamese, Zulu, and Tamil corpora[1] to expose our model to as diverse a set of languages as possible (in terms of morphology, phonology, language family, etc., in line with the Babel program goals).

The Limited LP includes a 10 hour training set (audio and transcripts) which we use for building acoustic and language models. We also estimate the parameters for our topic model from the same training data. The Babel corpora contain spontaneous conversational telephone speech, but without the constrained topic prompts of LDC's Fisher collections we would expect a sparse collection of topics. Yet for retrieval we are nonetheless able to leverage the information.

We estimate parameters $\phi^{(t)}$, $\kappa^{(d)}$, α, β, and ν on the training transcripts in each language, then use these parameters to infer $\theta^{(d)}$ (topic proportions) and $\kappa^{(d)}$ (cache usage) for each document in the held-out set. We use the inferred $\kappa^{(d)}$ and $\theta^{(d)}$ to perform the language model interpolation (Eqns. 3, 4). But also, the mean of the inferred $\kappa^{(d)}$ values for a corpus ought to provide a snapshot of the amount of repetition within.

Two trends emerge when we examine the mean over $\kappa^{(d)}$ by language. First, as shown in Table 1, the more latent topics are used, the lower the inferred κ values. Regardless of the absolute value, we see that κ for Vietnamese is consistently higher than the other languages. This fits our intuition about the languages given that the Vietnamese transcripts had syllable-level word units and we would expect to see more repetition.

[1]Releases babel106b-v0.2g, babel107b-v0.7, babel206b-v0.1e, and babel204b-v1.1b, respectively

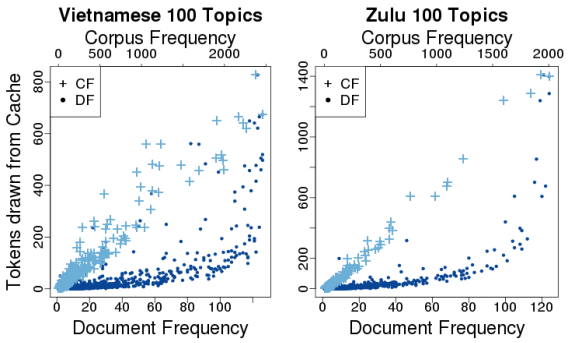

Figure 2: Cache and corpus frequencies for each word type in Vietnamese and Zulu training corpora.

Secondly we consider *which* words are drawn from the cache versus the topics during the inference process. Examining the final sampling state, we count how often each word in the vocabulary is drawn from the cache (where $k_{d,i} = 1$). Intuitively, this count is highly correlated ($\rho > 0.95$) with the corpus frequency of each word (cf. Figure 2). That is, cache states are assigned to word types most likely to repeat.

5.1 Perplexity

While our measurements of cache usage corresponds to intuition, our primary goal is to construct useful language models. After estimating parameters on the training corpora, we infer $\kappa^{(d)}$ and $\theta^{(d)}$ then measure perplexity using document-specific language models on the development set.

We compute perplexity on the topic unigram mixtures according to P_d and P_{dc} (Eqns.1 & 3). Here we do not interpolate with the base N-gram LM, so as to compare only unigram mixtures. Table 2 gives the perplexity for standard LDA (P_d only) and for our model with and without the cache added (κLDA$'$ and κLDA respectively).

With respect to perplexity, interpolating with the cache (κLDA) provides a significant boost in perplexity for all languages and values of $\mathcal{T}$. In general,

Language	$\mathcal{T}$	LDA	κLDA$'$	κLDA
Tagalog	50	142.90	163.30	**134.43**
	100	136.63	153.99	**132.35**
	150	139.76	146.08	**130.47**
	200	128.05	141.12	**129.94**
Vietnamese	50	257.94	283.52	**217.30**
	100	243.51	263.03	**210.05**
	150	232.60	245.75	**205.59**
	200	223.82	234.44	**204.25**
Zulu	50	**183.53**	251.52	203.56
	100	**179.44**	267.42	217.11
	150	**174.79**	269.01	223.90
	200	**175.65**	252.03	217.89
Tamil	50	**273.08**	356.40	283.82
	100	**265.02**	369.18	297.68
	150	**259.42**	361.79	301.92
	200	**236.30**	341.32	298.26

Table 2: Perplexities of topic unigram mixtures on held-out data, with and without cache.

perplexity decreases as the number of latent topics increases, excepting certain Zulu and Tamil models. For Tagalog and Vietnamese our cache-augmented model outperforms standard LDA model in terms of perplexity. However, as we will see in the next section, the lowest perplexity models are not necessarily the best in terms of retrieval performance.

6 Speech Retrieval

We evaluate the utility of our topic language model for speech retrieval via the term detection, or keyword search (KWS) task. Term detection accuracy is the primary evaluation metric for the Babel program. We use the topic and cache-augmented language models (Eqn. 4) to improve the speech recognition stage of the term detection pipeline, increasing overall search accuracy by 0.5 to 1.7% absolute over a typical N-gram language model.

The term detection task is this: given a corpus of audio documents and a list of terms (words or phrases), locate all occurrences of the key terms in the audio. The resulting list of detections is scored using Term Weighted Value (TWV) metric. TWV is a cost-value trade-off between the miss probability, $P(miss)$, and false alarm probability, $P(FA)$, averaged over all keywords (NIST, 2006). For comparison with previously published results, we score against the IARPA-supplied evaluation keywords.

We train acoustic and language models (LMs) on the 10 hour training set using the Kaldi toolkit (Povey and others, 2011), according to the training recipe described in detail by Trmal et al. (2014). While Kaldi produces different flavors of acoustic models, we report results using the hybrid HMM-DNN (deep neural net) acoustic models, trained with a minimum phone error (MPE) criterion, and based on PLP (perceptual linear prediction) features augmented with pitch. All results use 3-gram LMs with Good-Turing (Tagalog, Zulu, Tamil) or Modified Kneser-Ney (Vietnamese) smoothing. This AM/LM combination (our baseline) has consistently demonstrated state-of-the art performance for a single system on the Babel task.

As described, we estimate our model parameters $\phi^{(t)}$, $\kappa^{(d)}$, α, β, and ν from the training transcripts. We decode the development corpus with the baseline models, then infer $\theta^{(d)}$ and $\kappa^{(d)}$ from the first pass output. In principle we simply compute P_{Ldc} for each document and re-score the first pass output, then search for keywords.

Practical considerations for cache language models are, for example, just how big should the cache be, or should it decay, where words further away from the current word are discounted proportionally. In the Kaldi framework, speech is processed in segments (i.e. conversation turns). Current tools do not allow one to vary the language model within a particular segment (dynamically). With that in mind, our KWS experiments construct a different language model (P_{Ldc}) for each segment, where P_c is computed from all other segments in the current document except that being processed.

6.1 Results

We can show, by re-scoring LCVSR output with a cache-augmented topic LM, that both the document-specific topic (P_d) and cache (P_c) information together improve our overall KWS performance in each language, up to 1.7% absolute.

Figure 3 illustrates search accuracy (TWV) for each language under various settings for $\mathcal{T}$. It also captures alternatives to using $\kappa^{(d)}$ as an interpolation weight for the cached unigrams. To illustrate this contrast we substituted the training mean κ_{train} instead of $\kappa^{(d)}$ as the interpolation weight when computing P_{Ldc} (Eqn 4). Except for Zulu, the inferred

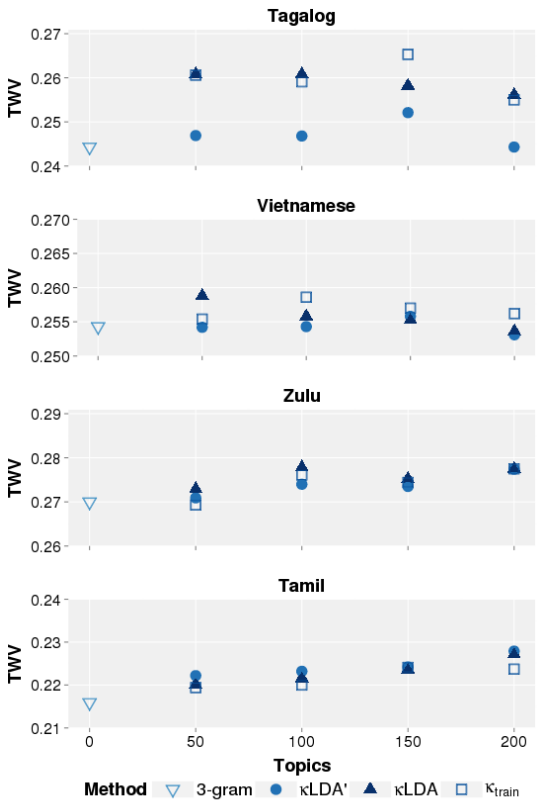

Figure 3: KWS accuracy for different choices of $\mathcal{T}$

$\kappa^{(d)}$ were more effective, but not hugely so.

The effect of latent topics $\mathcal{T}$ on search accuracy also varies depending on language, as does the overall effect of incorporating the cache in addition to latent topics (κLDA$'$ vs. κLDA). For example, in Tagalog, we observe most of the improvement over N-grams from the cache information, whereas in Tamil, the cache provided no additional information over latent topics.

The search accuracy for the best systems from Figure 3 are shown in Table 3 with corresponding choice of $\mathcal{T}$. Effects on WER was mixed under the cache model, improving Zulu from 67.8 to 67.6% and degrading Tagalog from 60.8 to 61.1%, with Vietnamese and Tamil unchanged.

7 Conclusions and Future Work

With our initial effort in formulating model combining latent topics with a cache-based language model, we believe we have presented a model that estimates both informative and useful parameters from the data and supports improved speech retrieval performance. The results presented here reinforce the conclusion that topics and repetition, *broad* and *local* context, are complementary sources of information for speech language modeling tasks.

Language	$\mathcal{T}$	3-gram	κLDA$'$	κLDA
Tagalog	50	0.244	0.247	**0.261**
Vietnamese	50	0.254	0.254	**0.259**
Zulu	100	0.270	0.274	**0.278**
Tamil	200	0.216	**0.228**	0.227

Table 3: Best KWS accuracy (TWV) is each language.

We hope to address two particular limitations of our model in the near future. First, all of our improvements are obtained adding unigram probabilities to a 3-gram language model. We would naturally want to extend our model to explicitly capture the cache and topic behavior of N-grams.

Secondly, our models are restricted by the first pass output of the LVCSR system. Keywords not present in the first pass cannot be recalled by a re-scoring only approach. An alternative would be to use our model to re-decode the audio and realize subsequently larger gains. Given that our re-scoring model worked sufficiently well across four fundamentally different languages, we are optimistic this would be the case.

Acknowledgements

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

We would also like to thank all of the reviewers for their insightful and helpful comments, and above all their time.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. In *JMLR*, volume 3, pages 993–1022. JMLR.org.

Chaitanya Chemudugunta, Padhraic Smyth, and Steyvers Mark. 2007. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 241. Mit Press.

Berlin Chen. 2009. Latent Topic Modelling of Word Co-occurence Information for Spoken Document Retrieval. In *Proc. of ICASSP*, pages 3961–3964. IEEE.

Jen-Tzung Chien and Chuang-Hua Chueh. 2011. Dirichlet Class Language Models for Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):482–495.

Justin Chiu and Alexander Rudnicky. 2013. Using conversational word bursts in spoken term detection. In *Proc. of Interspeech*, pages 2247–2251. ISCA.

Kenneth Ward Church and William A Gale. 1995. Poisson Mixtures. *Natural Language Engineering*, 1(2):163–190.

Philip R Clarkson and Anthony J Robinson. 1997. Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache. In *Proc. of ICASSP*, volume 2, pages 799–802. IEEE.

Radu Florian and David Yarowsky. 1999. Dynamic Nonlocal Language Modeling via Hierarchical Topic-based Adaptation. In *Proc. of ACL*, pages 167–174. ACL.

Mary Harper. 2011. Babel BAA. `http://www.iarpa.gov/index.php/research-programs/babel/baa`.

Marti A Hearst and Christian Plaunt. 1993. Subtopic Structuring for Full-length Document Access. In *Proc. of SIGIR*, pages 59–68. ACM.

Aaron Heidel, Hung-an Chang, and Lin-shan Lee. 2007. Language Model Adaptation Using Latent Dirichlet Allocation and an Efficient Topic Inference Algorithm. In *Proc. of Interspeech*. ICSA.

Thomas Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1):177–196.

Houghton-Mifflin. 1997. *The American Heritage College Dictionary*. Houghton Mifflin.

Bo-June Paul Hsu and James Glass. 2006. Style & Topic Language Model Adaptation Using HMM-LDA. In *Proc. of EMNLP*, pages 373–381. ACL.

Songfang Huang and Steve Renals. 2008. Unsupervised Language Model Adaptation Based on Topic and Role Information in Multiparty Meetings. In *Proc. of Interspeech*. ICSA.

Frederick Jelinek, Bernard Merialdo, Salim Roukos, and Martin Strauss. 1991. A Dynamic Language Model for Speech Recognition. *HLT*, 91:293–295.

Sanjeev Khudanpur and Jun Wu. 1999. A Maximum Entropy Language Model Integrating N-grams and Topic Dependencies for Conversational Speech Recognition. In *Proc. of ICASSP*, volume 1, pages 553–556. IEEE.

Roland Kuhn and Renato De Mori. 1990. A Cache-based Natural Language Model for Speech Recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.

Xiaoyong Liu and W Bruce Croft. 2004. Cluster-based Retrieval Using Language Models. In *Proc. of SIGIR*, pages 186–193. ACM.

Yang Liu and Feifan Liu. 2008. Unsupervised Language Model Adaptation via Topic Modeling Based on Named Entity Hypotheses. In *Proc. of ICASSP*, pages 4921–4924. IEEE.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. In *Proc. of Interspeech*. ICSA.

David Miller et al. 2007. Rapid and Accurate Spoken Term Detection. In *Proc. of Interspeech*. ICSA.

NIST. 2006. The Spoken Term Detection (STD) 2006 Evaluation Plan. `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`. [Online; accessed 28-Feb-2013].

Barbara et al. Peskin. 1996. Improvements in Switchboard Recognition and Topic Identification. In *Proc. of ICASSP*, volume 1, pages 303–306. IEEE.

Daniel Povey et al. 2011. The Kaldi Speech Recognition Toolkit. In *Proc. of ASRU Workshop*. IEEE.

Ronald Rosenfeld. 1994. *Adaptive Statistical Language Modeling: a Maximum Entropy Approach*. Ph.D. thesis, CMU.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic Topic Models. *Handbook of Latent Semantic Analysis*, 427(7):424–440.

Jan et al. Trmal. 2014. A Keyword Search System Using Open Source Software. In *Proc. of Spoken Language Technoloty Workshop*. IEEE.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why Priors Matter. In *Proc. of NIPS*, volume 22, pages 1973–1981. NIPS.

Xing Wei and W Bruce Croft. 2006. LDA-based Document Models for Ad-hoc Retrieval. In *Proc. of SIGIR*, pages 178–185. ACM.

Jonathan Wintrode and Khudanpur Sanjeev. 2014. Combining Local and Broad Topic Context to Improve Term Detection. In *Proc. of Spoken Language Technoloty Workshop*. IEEE.

Jonathan Wintrode. 2014a. Can you Repeat that? Using Word Repetition to Improve Spoken Term Detection. In *Proc. of ACL*. ACL.

Jonathan Wintrode. 2014b. Limited Resource Term Detection For Effective Topic Identification of Speech. In *Proc. of ICASSP*. IEEE.

YouTube. 2015. Statistics - YouTube. `http://www.youtube.com/yt/press/statistics.html`, February.