An LDA-based Topic Selection Approach to Language Model Adaptation for Handwritten Text Recognition

Jafar Tanha, Jesse de Does and Katrien Depuydt
Institute for Dutch Lexicology (INL)
jafar.tanha.pnu@gmail.com, {jesse.dedoes, katrien.depuydt}@inl.nl

Abstract

Typically, only a very limited amount of in-domain data is available for training the language model component of an Handwritten Text Recognition (HTR) system for historical data. One has to rely on a combination of in-domain and out-of-domain data to develop language models. Accordingly, domain adaptation is a central issue in language modeling for HTR. We pursue a topic modeling approach to handle this issue, and propose two algorithms based on this approach. The first algorithm relies on posterior inference for topic modeling to construct a language model adapted to the development set, and the second algorithm proceeds by iterative selection, using a new ranking criterion, of topic-dependent language models. Our experimental results show that both approaches clearly outperform a strong baseline method.

1 Introduction

Huge amounts of handwritten historical documents are nowadays being published by on-line digital libraries as document images. The content of these documents is of great interest to historians, linguists and literary scholars alike. However, if the transcription of the documents is not available for information retrieval, we can hardly consider this content to be accessible for research. Full manual transcription is slow and costly, but the development of efficient and cost-effective approaches for the indexing, search and full transcription of historical handwritten document images can benefit from modern Handwritten Text Recognition (HTR) technology (Sánchez et al., 2013).

An indispensable component of state-of-the-art HTR is language modeling (Plötz and Fink, 2009) (Espana-Boquera et al., 2011), which is necessary to guide the decoding process by ranking and constraining the possible recognition hypotheses. Language modeling has proven extremely successful in improving results of Automatic Speech Recognition (Chelba et al., 2012), which is a very similar task from the technical point of view. Highly effective language models in this field have been developed from huge language corpora. cf. for instance (Chelba et al., 2012). Language models are usually constructed from large text corpora which – ideally – are *in-domain*, linguistically close to the language of the document collection which is being processed. However, for HTR of historical documents, obtaining effective models is much less straightforward: models built from the strictly in-domain data are generally unsatisfactory because not enough data can be obtained to avoid overfitting, and in order to exploit the larger pool of out-domain data one has to surmount two difficulties: (1) indiscriminate use of *out-of-domain* data may not benefit, in fact even deteriorate system performance and (2) the use of the complete out-domain data for training may increase the complexity of the system, making the decoding process almost untractable (Axelrod et al., 2011; Tanha et al., 2014).

The above-mentioned issues are typically dealt with by using *domain adaptation* techniques (Axelrod et al., 2011) (Foster et al., 2010) (Jiang and Zhai, 2007), which aim to leverage the knowledge that can be obtained from the out-of-domain data by tuning it to the in-domain data.

In this paper, we study the application of topic modeling-based approaches to the task of improving the language modeling component of the HTR system by domain adaptation. Our approach is characterized by the combination of the topic modeling approach with *intelligent sample selection* methods. We first propose a Latent Dirichlet Allocation (LDA)-based language model adap-

Proceedings of Recent Advances in Natural Language Processing, pages 646–653,
Hissar, Bulgaria, Sep 7–9 2015.

tation framework (Blei et al., 2003). We then develop an algorithm for language model adaptation using the result of topic modeling and a new language model ranking criterion to select the most relevant topics. In our experiments, we use the TRANSCRIPTORIUM HTR engine described in (Sánchez et al., 2013) on a set of digitised images of manuscripts written by the 18th and early 19th-century British philosopher Jeremy Bentham[1]. We show that our techniques improve the performance of the HTR system. Besides producing an adapted language model, the proposed methods also reduce the computational resources needed to exploit a large amount of out-domain data in the decoding process of the HTR system.

The rest of the paper is organised as follows. We refer the reader to related work in section 2. Our approaches to sample selection are described in detail in section 3 and evaluated in section 4. Results are reported in section 5. Section 6 addresses the discussion and conclusion.

2 Related Work

Statistical language models assign probabilities to sequences of words. Typically, the probability of a word is estimated on the basis of a limited history, consisting of some fixed number n of preceding words. This has the drawback that long-range dependencies cannot be exploited. Several approaches have been proposed to overcome this problem, such as Cache-based (Kuhn and De Mori, 1990) or Trigger-based (Lau et al., 1993) language models.

Taking into account that a language model built for domain-specific data can give low perplexity, topic modeling can be a promising approach for language model adaptation. A language model training corpus may contain many topics. As a result, the corpus can be divided into topic-specific subcorpora. The distribution of topics in the corpus may be determined manually, or by automatic, unsupervised techniques. A practical approach to language modeling will have to rely on the latter approach.

The leading paradigm in unsupervised topic modeling is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA and similar approaches can be used for the language model adaptation problem. There are several studies for language model adaptation using LDA models. In (Liu and Liu, 2008) a new mixture topic model is proposed for LDA based language model adaptation. Hsu *et al.* (2006) proposed a method for adaptation using hidden Markov model with LDA model. In (Eidelman et al., 2012) LDA model is used to compute topic-dependent lexical weighting probabilities for domain adaptation. Iyer *et al.* (1996) used a clustering approach to build topic clusters for language model adaptation. In (Bellegarda, 2000) Latent Semantic Analysis is applied to map documents into a topic space for language model adaptation. Gildea *et al.* (1999) proposed a language model adaptation approach using the probabilistic extension of LSA (pLSA). In (Tam and Schultz, 2005), an LDA model is applied to language model adaptation. This method interpolates the background language model with the dynamic unigram language model generated by the LDA model. Heidel *et al.* (2007) applied an LDA-based topic inference approach to language model adaptation.

As mentioned in the introduction, the main characteristic of our approach is that we use topic modeling in conjunction with *Intelligent sample selection* techniques. Unlike current approaches, like (Liu and Liu, 2008) (Eidelman et al., 2012), which use all documents of each topic for adaptation, we select the most relevant resources. In this way, language model adaptation yields a model that matches better to the domain, but also reduces the computational complexity of the HTR system by producing more compact language models.

More specifically, we propose an iterative approach to language model adaptation using LDA modeling. Since perplexity does not always correlate well to the recognition accuracy of the HTR system, we use a new criterion for related topic selection using the combination of the perplexity and the size of out of vocabulary of the documents. We then use a topic mixture approach for language model adaptation.

3 Topic Modeling for Language Model Adaptation

We first briefly review the Latent Dirichlet Allocation (LDA) framework. We then formulate the problem and introduce the proposed methods to language model adaption for improving the performance of the HTR system.

[1] Images and transcriptions have been produced in the *Transcribe Bentham* project (Moyle et al., 2011), http://www.ucl.ac.uk/transcribe-bentham

3.1 LDA Models and Language Model Adaptation

The frequency distribution of words in text is highly dependent on the "topic" of the text. A topic model captures this intuition in a mathematical framework, and allows discovering a set of topics from a collection of documents. Blei *et al.* (2003) introduced a new approach, Latent Dirichlet Allocation (LDA). LDA is a generative approach characterized by the topic-word distribution ϕ and the topic distribution θ for each document. This method imposes a Dirichlet distribution on the topic mixture weights corresponding to the documents in the corpus. Figure 1 shows the graphical representation of the LDA model, where α is the parameter of the Dirichlet prior on the per-document topic distribution, D is the number of documents, β is the parameter of the Dirichlet prior on the per-topic word distribution, θ_d is the topic distribution for document d, N is the number of words in document d $z_{d,n}$ is the topic for the d^{th} word in document n, and $w_{d,n}$ is the specific word.

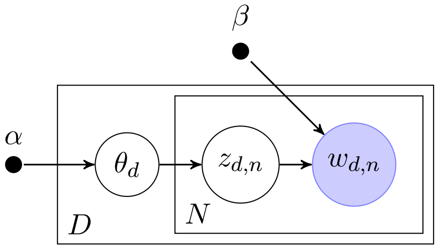

Figure 1: The graphical representation of LDA.

In order to apply topic-based language modeling, we need to be able to determine the topic distribution for an unseen document (topic inference). We use the collapsed variational inference method (CVB) for LDA (Mukherjee and Blei, 2009) in our experiments.

3.2 Problem Formulation

In this paper, we use topic modeling to identify relevant resources for language model adaptation. We assume a partition ($\mathcal{B}_0$, $\mathcal{B}_1$) of the in-domain corpus $\mathcal{B}$, see Figure 2. In the setting of handwritten text recognition, $\mathcal{B}_0$ could for instance be the HTR training set or some other portion of a transcribed corpus, and $\mathcal{B}_1$ is the rest of the text. We then use $\mathcal{E}$ corpus as a general large out-of-domain corpus. Our goal here is to find an informative subset $\mathcal{E}_1$ of resources from the $\mathcal{E}$ corpus, which is relevant to the $\mathcal{B}_0$ collection, and to exploit this for domain adaptation.

The adapted language model can be then obtained as follows: for a word sequence W, let

$$P(W) = \lambda_{\mathcal{B}_0} P_{\mathcal{B}_0}(W) + \lambda_{\mathcal{B}_1} P_{\mathcal{B}_1}(W) + \lambda_{\mathcal{E}_1} P_{\mathcal{E}_1}(W) \tag{1}$$

where $\lambda_{\mathcal{B}_0}$, $\lambda_{\mathcal{B}_1}$ and $\lambda_{\mathcal{E}_1}$ are interpolation weights. We use the *SRILM* toolkit (Stolcke et al., 2011) to find the optimal values for the weights in equation (1).

In (1) the third term is the resulting adapted language model, which we formulate as:

$$P_{\mathcal{E}_1}(W) = \Sigma_{i=1}^{K'} \gamma_i P_{z_j}(w_i | w_{i-1}^{i-n+1}) \tag{2}$$

where γ_i is the mixture weight and K' is the number of relevant topics to the domain ($K' \ll K$). Based on this formulation, we propose two algorithms to handle the equation (2).

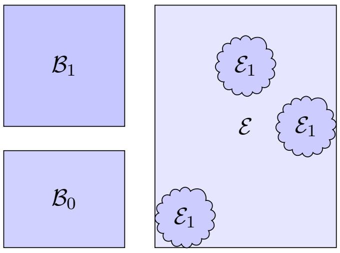

Figure 2: Resource situation consisting of in-domain corpora $\mathcal{B}_0$ and $\mathcal{B}_1$ and out-of-domain corpus $\mathcal{E}$. The aim of intelligent sample selection is to pick out the informative bits $\mathcal{E}_1$ from $\mathcal{E}$.

3.3 LDA Inference for LM Adaptation

We first introduce an unsupervised language model adaptation method using posterior inference for LDA, which we call *Inference-Based Topic Selection*. In accordance with the sample selection approach, the goal is to pick the most relevant documents to in-domain data from the out-of-domain corpus. We start by applying LDA to the $\mathcal{E}$ corpus to construct a model with N topics, and we select, for each topic found, the set of most relevant (high-confidence) documents from $\mathcal{E}$. Next, the topic model is used to inference the topic distribution of the development set ($\mathcal{B}_0$). Finally, based on the distribution found in the development set, the algorithm selects the most relevant topics. The sets of all high-confidence documents from the selected topics are then used to train language models, and the interpolation of the resulting language models will be the output of the proposed algorithm. The pseudo-code of the Inference-Based Topic-Selection algorithm is presented in Algorithm 1. In the experiment section, we will describe the tuning parameters of the algorithm for improving the HTR system.

Algorithm 1 Inference-Based Topic-Selection

```
Initialize:
$\mathcal{E}$: out-domain data; $conf \leftarrow 0$; // A pre-defined threshold for confidence measure;
$\theta \leftarrow$ Threshold for Topic Selection; $N \leftarrow$ Identify maximum number of topics;
$TopicModel \leftarrow \{$Make a topic model with $N$ topics for $\mathcal{E}$ resources using LDA model $\}$;
for each $T_i \in TopicModel$ do
    if ( probability of $d_j \in T_i$ is greater than $conf$ ) then
        $D_i \leftarrow D_i + d_j$;
    end if
end for
$DocSet \leftarrow \{D_i \mid D_i$ is more relevant document for topic $T_i\}$
$DevelopmentSetTopics \leftarrow$ Infer Topics for Development set using the resulting $TopicModel$;
for each $t_i \in DevelopmentSetTopics$ do
    if probability of $t_i$ is greater than $\theta$ then
        $BestTopicSet \leftarrow BestTopicSet + \{t_i \in TopicModel\}$;
    end if
end for
Build $LM_i$ for each $t_i \in BestTopicSet$;
$InterpolatedLMs \leftarrow \sum_i \lambda_i LM_i$;
Output:
    $InterpolatedLMs$ and $BestTopicSet$;
```

3.4 Iterative Topic Selection for Language Model Adaptation

We present an iterative algorithm, *Iterative Topic-Selection*, for topic selection based on a new ranking criterion. As described in Section 3.3, first the algorithm builds a topic model for the $\mathcal{E}$ corpus. Then for each topic, we construct a language model. The resulting language models are evaluated against the development dataset. Since comparing and ranking different resources by the usual perplexity-related criteria alone is much less appropriate (Moore and Lewis, 2010; Axelrod et al., 2011; Tanha et al., 2014), we use a new criterion for related topic selection in terms of the perplexity and out-of-vocabulary (OOV) word rate in the following section. Note that, as mentioned in Section 3.3, we do not use all topics for language model adaptation, but only the most relevant ones.

Next, the algorithm ranks the resulting language model of each topic using the new criterion. The language models of the related topics are then interpolated until some stopping condition is reached. Algorithm 2 shows the pseudo-code of the proposed algorithm. Finally, the interpolated language model is returned as the adapted language model, which can be used in (1).

Algorithm 2 Iterative Topic-Selection

```
Initialize:
$\mathcal{E}$: out-domain data; $conf \leftarrow 0$; // A pre-defined threshold for confidence measure;
$\theta \leftarrow$ Threshold for Topic Selection; $N \leftarrow$ Identify maximum number of topics;
$TopicModel \leftarrow \{$Make a topic model with $N$ topics for $\mathcal{E}$ resources using LDA model $\}$;
for each $T_i \in TopicModel$ do
    if ( probability of $d_j \in T_i$ is greater than $conf$ ) then
        $D_i \leftarrow D_i + d_j$;
    end if
end for
$DocSet \leftarrow \{D_i \mid D_i$ is more relevant document for topic $T_i\}$
for each $D_i \in DocSet$ do
    $LM_i \leftarrow$ Train a Language Model for $D_i$;
    $EvalSet_i \leftarrow$ Evaluate $LM_i$ using a development set;
    $RankSet_i \leftarrow$ Assign ranks to the evaluated sets using (3);
end for
$BestRank \leftarrow$ Find the best rank based on the ranking criterion;
$BestTopicSet \leftarrow$ Find the best topic based on the ranking criterion;
while ( $TopicSet$ ) do
    Interpolate the related LMs $T_i$ and $T_{BestRank}$, where $T_i$ is the second best related topic;
    $New_{rank} \leftarrow$ Compute new rank for the interpolated LM;
    if ($New_{rank} < BestRank$) then
        $BestTopicSet \leftarrow BestTopicSet + T_i$;
        $BestRank \leftarrow New_{rank}$;
    else
        Break;
    end if
end while
Output:
    Adapted Language model and $RelatedTopics$;
```

3.5 Ranking based on Out-of-vocabulary and Perplexity

Algorithm 2 first builds a language model for each topic, and subsequently assumes that the resulting language models can to be ranked in an appropriate way. Current approaches to rank language models use perplexity as a criterion (Moore and Lewis, 2010). However, perplexity as a criterion is unreliable when the text contains more than a small portion of OOV words (Tanha et al., 2014).

Let $|\mathcal{V}|$ be the number of running words (i.e. tokens) of in the evaluation data, $|OOV|$ be the number of running out-of-vocabulary words, and PPL denote the perplexity. We use the following ranking function combining OOV rate and perplexity:

$$Rank(LM_i) = \log PPL \times \frac{|OOV|}{|\mathcal{V}|} \quad (3)$$

We apply this *Multiplicative* ranking function to rank resources for sample selection in algorithm 2.

4 Experiments

In this section we perform several experiments on linguistic resources to show the effect of the proposed methods for language model adaption on the HTR system. In order to evaluate the proposed methods, it is important to compare them to a strong baseline, in our case a well-tuned linear interpolation of in-domain and out-of-domain language models.

4.1 Dataset

We make use of the English-language data processed in the TRANSCRIPTORIUM (Sánchez et al., 2013) project for the evaluation of HTR performance. This collection consists of a set of images and with ground truth transcriptions of Bentham manuscripts. Part of the ground truth transcriptions is used for language modeling, a held-out set is used for testing HTR. In addition to this, we use the corpus of all transcribed Bentham manuscripts (about 15.000 pages and 5m words), as obtained from the *Transcribe Bentham* project (Moyle et al., 2011), and the public part of the ECCO (*Eighteenth Century Collections Online*[2]), about 70m words.

With these two corpora, we make a two-level in-domain/out-domain distinction: The ECCO corpus is considered as a general out-of-domain resource. Within the set of Bentham transcriptions, we distinguish the set of *Batch 1 ground truth transcriptions* as an in-domain resource and the rest of the available transcriptions as Bentham out-of-domain.

In the experiments, we use a separate development set for tuning parameters of the proposed methods and a separate test set for evaluating the HTR system, consisting of the held-out data from the "Batch 1" set.

4.2 Experimental Setup

We perform the following experiments to evaluate the baseline methods and the proposed Inference-Based Topic-Selection and Iterative Topic-Selection algorithms.

Baseline: Language model interpolation Our first set of experiments is about finding an optimal way to combine in-domain and out-of domain resources by language model interpolation. We explore the effect of tuning language model interpolation parameters and HTR dictionary selection settings on the performance of the HTR system. We have applied the following scenarios in our experiments:

1. Combining two Bentham resources ($\mathcal{B}_0$ and $\mathcal{B}_1$) and using a dictionary from the merged data to train the language model (*Merged-InOut-Dic-InOut*).
2. Interpolating Bentham $\mathcal{B}_0$ and $\mathcal{B}_1$ resources using a HTR dictionary from both $\mathcal{B}_0$ and $\mathcal{B}_1$ domain data (*Inter-InOut-Dic-InOut*).
3. Interpolating Bentham $\mathcal{B}_0$ and $\mathcal{B}_1$ resources with the ECCO collection using the dictionary from Bentham $\mathcal{B}_0$ and $\mathcal{B}_1$ data (*Inter-InOutECCO-Dic-InOut*).
4. Interpolating the Bentham $\mathcal{B}_0$ and $\mathcal{B}_1$ resources with ECCO collection using dictionary from all of them (*Inter-InOutECCO-Dic-InOutECCO*).

Inference-Based Topic-Selection We perform several experiments with different numbers of topics and different values for the θ threshold.

Iterative Topic-Selection The following scenarios are used for the Iterative Topic-Selection algorithm:

Single Iteration (Best Topic): In this scenario, a single iteration of the algorithm is used to select the most relevant topic. The selected set is then used to build a language model. We vary the number of topics and the threshold for document selection.

Multiple Iterations: In this scenario, the algorithms perform several iterations. At each iteration the resulting best-fitted language model is interpolated with the last language model.

5 Results

We have considered three main evaluation criteria for each experiment, the general word error rate (WER), the word error rate without taking the first word of each line into account[3], and the character error rate (CER).

[2] http://www.textcreationpartnership.org/tcp-ecco/

[3] Current HTR is line-based, which means that language modeling fails at the line boundaries, most notably for hyphenated words.

Method	WER %	WER without first word	CER %	OOV %	Size of model (1-grams,2-grams)
Initial model using only Batch 1 training set	34.5	34.3	19.9	9.44	(1894 , 6641)
Merged-InOut-Dic-InOut	34.01	-	-	-	(13211 , 808724)
Inter-InOut-Dic-InOut	30.02	24.57	-	-	(12966, 795029)
Inter-In+OutECCO-Dic-InOutECCO	31.7	26	16.5	-	(12966 , 2817124)
Inter-InOutECCO-Dic-InOut	30.7	25.3	15.9	-	(12966, 2833622)
Inter-InOutECCO-Dic-InOutECCO	**28.3**	22.7	14.7	5.4	**(64416, 5811657)**

Table 1: The results of the baseline methods for HTR system

Number of Topics			WER%	WER without first word	CER%	Size of model (1-grams,2-grams)
#Topics	Threshold for document selection	θ				
30	0.3	0.2	27.4	22.0	14.4	(51682, 1281779)
30	0.8	0.2	27.4	22.0	-	(40010, 984660)
40	0.4	0.1	**27.3**	22.0	14.4	**(55054, 1073172)**
50	0.7	0.2	27.5	22.1	-	(41091, 996482)
70	0.2	0.2	27.4	22.1	-	(41476, 1005868)
70	0.2	0.1	27.4	22.1	-	(49755, 132526)
100	0.2	0.1	27.3	22.0	-	(56533, 1312630)

Table 2: The results of Inference Best-Topic approach

Number of Topics		WER %	WER without first word	CER%	Size of model (1-grams,2-grams)
#Topics	Threshold for document selection				
10	0.5	27.2	21.9	14.3	(63379, 1595014)
10	0.8	27.2	21.7	-	(47839, 1345021)
20	0.3	27.5	22.1	-	(54626, 1355764)
40	0.3	27.3	21.9	-	(46880, 1164894)
40	0.4	**27.2**	21.8	14.3	**(46227, 1180713)**
50	0.3	27.3	22.0	-	(49517, 1286652)
50	0.4	27.3	21.9	-	(53761, 1187976)
100	0.3	27.6	22.2	-	(51512, 1457114)

Table 3: The results of the Best-Topic approach

Number of Topics		WER %	WER without first word	CER%	Size of model
#Topics	Threshold for document selection				
10	0.9	27.3	22.0	14.4	(64034, 1439691)
15	0.9	27.8	22.5	-	(53159, 1605800)
20	0.9	27.3	22.0	-	(63113, 1355451)
30	0.9	27.4	22.0	-	(58891, 1235159)
40	0.9	27.4	22.0	-	(48768, 1078245)
50	0.5	**27.2**	21.9	14.3	**(61682, 1567515)**
50	0.9	27.3	21.9	-	(56580, 1100616)
70	0.8	27.4	22.0	-	(52146, 1154698)
100	0.3	27.6	22.2	-	(65343, 1455298)

Table 4: The results of the Iterative Best-Topic approach

In the first experiment we also include the size of OOV sets. In each table the best results have been boldfaced.

Table 1 shows the results of interpolating the language model from Bentham in-domain data with the language models from the Bentham out-of-domain and ECCO resources. This procedure improves the performance of the HTR system by 6.2%. In other words, these results emphasize that the out-of-domain data contains useful information.

Table 2 shows the performance of the HTR system using the proposed Inference Best-Topic algorithm. In Table 2 the first column shows the number of topics identified. The second and third columns are the threshold for document selection for each topic and the threshold for the related topic selection respectively. The Inference Best-Topic algorithm performs better than the baseline methods in most of the cases. Furthermore, the resulting language model is much more compact than the baseline model.

We continue with the experiments for the *Iterative Topic-Selection* algorithm. In the first experiment, the *Iterative Topic-Selection* algorithm (single iteration) finds the most relevant language model for adaptation. Table 3 shows the results of this experiment.

The Iterative Topic-Selection algorithm (multiple iterations) deploys the interpolation of the most relevant language models. The results have been reported in Table 4. The results of both experiments emphasize that the proposed methods for language model adaptation outperform the baseline and produce a more domain-specific language model.

6 Conclusion

We have studied and tested several ways in which domain adapted language modeling can improve hand-written text recognition results, when the resulting language models are deployed in the TRANSCRIPTORIUM HTR system. Our methods for the combination of in-domain and out-of domain data have been shown to yield improvement in HTR results, both using established techniques (model interpolation) and novel approaches for language model adaptation. Consistent to our hypothesis, the proposed methods outperform the baseline, both in terms of HTR accuracy and in terms of model complexity. The experimental results show that our proposed approaches for domain adaptation can effectively exploit informative data from the out of domain data and improve the recognition performance of the HTR system significantly.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600707 - tranScriptorium.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362. ACL.

J.R. Bellegarda. 2000. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296, Aug.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Ciprian Chelba, Dan Bikel, Maria Shugrina, Patrick Nguyen, and Shankar Kumar. 2012. Large scale language modeling in automatic speech recognition. Technical report, Google.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the ACL: Short Papers-Volume 2*, pages 115–119. ACL.

Salvador Espana-Boquera, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez. 2011. Improving offline handwritten text recognition with hybrid hmm/ann models. *PAMI, IEEE Transactions on*, 33(4):767–779.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on EMNLP*, pages 451–459.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *ACL*, volume 2007, page 22.

R. Kuhn and R. De Mori. 1990. A cache-based natural language model for speech recognition. *IEEE Trans. PAMI*, 12(6):570–583, June.

R. Lau, R. Rosenfeld, and S. Roukos. 1993. Trigger-based language models: A maximum entropy approach. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 45–48, Minneapolis MN.

Yang Liu and Feifan Liu. 2008. Unsupervised language model adaptation via topic modeling based on named entity hypotheses. In *Acoustics, Speech and Signal Processing, ICASSP*, pages 4921–4924. IEEE.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics.

Martin Moyle, Justin Tonra, and Valerie Wallace. 2011. Manuscript transcription by crowdsourcing: Transcribe bentham. *LIBER Quarterly*, 20(3).

Indraneel Mukherjee and David M Blei. 2009. Relative performance guarantees for approximate inference in latent dirichlet allocation. In *NIPS*, pages 1129–1136.

Thomas Plötz and Gernot A Fink. 2009. Markov models for offline handwriting recognition: a survey. *Journal on Document Analysis and Recognition*, 12(4):269–298.

Joan Andreu Sánchez, Günter Mühlberger, Basilis Gatos, Philip Schofield, Katrien Depuydt, Richard M Davis, Enrique Vidal, and Jesse de Does. 2013. transcriptorium: a european project on handwritten text recognition. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 227–228. ACM.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. Srilm at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, page 5.

Yik-Cheung Tam and Tanja Schultz. 2005. Dynamic language model adaptation using variational bayes inference. In *INTERSPEECH*, pages 5–8. Citeseer.

Jafar Tanha, Jesse de Does, and Katrien Depuydt. 2014. An intelligent sample selection approach to language model adaptation for hand-written text recognition. *Proceedings of the 2014 ICFHR conference*, pages 349–355.