Topic Adaptation for Lecture Translation through Bilingual Latent Semantic Models

Nick Ruiz[*]
Free University of Bozen-Bolzano
Bolzano, Italy
nicruiz@fbk.eu

Marcello Federico
FBK-irst
Fondazione Bruno Kessler
Trento, Italy
federico@fbk.eu

Abstract

This work presents a simplified approach to bilingual topic modeling for language model adaptation by combining text in the source and target language into very short documents and performing Probabilistic Latent Semantic Analysis (PLSA) during model training. During inference, documents containing only the source language can be used to infer a full topic-word distribution on all words in the target language's vocabulary, from which we perform Minimum Discrimination Information (MDI) adaptation on a background language model (LM). We apply our approach on the English-French IWSLT 2010 TED Talk exercise, and report a 15% reduction in perplexity and relative BLEU and NIST improvements of 3% and 2.4%, respectively over a baseline only using a 5-gram background LM over the entire translation task. Our topic modeling approach is simpler to construct than its counterparts.

1 Introduction

Adaptation is usually applied to reduce the performance drop of Statistical Machine Translation (SMT) systems when translating documents that deviate from training and tuning conditions. In this paper, we focus primarily on language model (LM) adaptation. In SMT, LMs are used to promote fluent translations. As probabilistic models of sequences of words, language models guide the selection and ordering of phrases in translation. With respect to LM training, LM adaptation for SMT tries to improve an existing LM by using smaller amounts of texts. When adaptation data represents the translation task domain one generally refers to *domain adaptation*, while when they just represent the content of the single document to be translated one typically refers to *topic adaptation*.

We propose a cross-language topic adaptation method, enabling the adaptation of a LM based on the topic distribution of the source document during translation. We train a latent semantic topic model on a collection of bilingual documents, in which each document contains both the source and target language. During inference, a latent topic distribution of words across both the source and target languages is inferred from a source document to be translated. After inference, we remove all source language words from the topic-word distributions and construct a unigram language model which is used to adapt our background LM via Minimum Discrimination Information (MDI) estimation (Federico, 1999, 2002; Kneser et al., 1997).

We organize the paper as follows: In Section 2, we discuss relevant previous work. In Section 3, we review topic modeling. In Section 4, we review MDI adaptation. In Section 5, we describe our new bilingual topic modeling based adaptation technique. In Section 6, we report adaptation experiments, followed by conclusions and future work in Section 7.

2 Previous work

Zhao et al. (2004) construct a baseline SMT system using a large background language model and use it to retrieve relevant documents from large monolin-

[*] This work was carried out during an internship period at Fondazione Bruno Kessler.

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 294–302,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

gual corpora and subsequently interpolate the resulting small domain-specific language model with the background language model. In Sethy et al. (2006), domain-specific language models are obtained by including only the sentences that are similar to the ones in the target domain via a relative entropy based criterion.

Researchers such as Foster and Kuhn (2007) and Koehn and Schroeder (2007) have investigated mixture model approaches to adaptation. Foster and Kuhn (2007) use a mixture model approach that involves splitting a training corpus into different components, training separate models on each component, and applying mixture weights as a function of the distances of each component to the source text. Koehn and Schroeder (2007) learn mixture weights for language models trained with in-domain and out-of-domain data respectively by minimizing the perplexity of a tuning (development) set and interpolating the models. Although the application of mixture models yields significant results, the number of mixture weights to learn grows linearly with the number of independent language models applied.

Most works focus on monolingual language model adaptation in the context of automatic speech recognition. Federico (2002) combines Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) for topic modeling with the minimum discrimination information (MDI) estimation criterion for speech recognition and notes an improvement in terms of perplexity and word error rate (WER). Latent Dirichlet Allocation (LDA) techniques have been proposed as an alternative to PLSA to construct purely generative models. LDA techniques include variational Bayes (Blei et al., 2003) and HMM-LDA (Hsu and Glass, 2006).

Recently, bilingual approaches to topic modeling have also been proposed. A Hidden Markov Bilingual Topic AdMixture (HM-BiTAM) model is proposed by Zhao and Xing (2008), which constructs a generative model in which words from a target language are sampled from a mixture of topics drawn from a Dirichlet distribution. Foreign words are sampled via alignment links from a first-order Markov process and a topic specific translation lexicon. While HM-BiTAM has been used for bilingual topic extraction and topic-specific lexicon mapping in the context of SMT, Zhao and Xing (2008) note that HM-BiTAM can generate unigram language models for both the source and target language and thus can be used for language model adaptation through MDI in a similar manner as outlined in Federico (2002). Another bilingual LSA approach is proposed by Tam et al. (2007), which consists of two hierarchical LDA models, constructed from parallel document corpora. A one-to-one correspondence between LDA models is enforced by learning the hyperparameters of the variational Dirichlet posteriors in one LDA model and bootstrapping the second model by fixing the hyperparameters. The technique is based on the assumption that the topic distributions of the source and target documents are identical. It is shown by Tam et al. (2007) that the bilingual LSA framework is also capable of adapting the translation model. Their work is extended in Tam and Schultz (2009) by constructing parallel document clusters formed by monolingual documents using M parallel seed documents.

Additionally, Gong et al. (2010) propose translation model adaptation via a monolingual LDA training. A monolingual LDA model is trained from either the source or target side of the training corpus and each phrase pair is assigned a phrase-topic distribution based on:

$$\hat{M_i^j} = \frac{w_k^j \cdot M_i^j}{\sum_{k=1}^{m} w_k^j}, \qquad (1)$$

where M^j is the topic distribution of document j and w_k is the number of occurrences of phrase pair X_k in document j.

Mimno et al. (2009) extend the original concept of LDA to support polylingual topic models (PLTM), both on parallel (such as EuroParl) and partly comparable documents (such as Wikipedia articles). Documents are grouped into tuples $\mathbf{w} = (\mathbf{w}^1, ..., \mathbf{w}^L)$ for each language $l = 1, ..., L$. Each document $\mathbf{w}^l$ in tuple $\mathbf{w}$ is assumed to have the same topic distribution, drawn from an asymmetric Dirichlet prior. Tuple-specific topic distributions are learned using LDA with distinct topic-word concentration parameters β^l. Mimno et al. (2009) show that PLTM sufficiently aligns topics in parallel corpora.

3 Topic Modeling

3.1 PLSA

The original idea of LSA is to map documents to a *latent semantic space*, which reduces the dimensionality by means of singular value decomposition (Deerwester et al., 1990). A word-document matrix A is decomposed by the formula $A = U\Sigma V^t$, where U and V are orthogonal matrices with unit-length columns and Σ is a diagonal matrix containing the singular values of A. LSA approximates Σ by casting all but the largest k singular values in Σ to zero.

PLSA is a statistical model based on the likelihood principle that incorporates mixing proportions of latent class variables (or topics) for each observation. In the context of topic modeling, the latent class variables $z \in Z = \{z_1, ..., z_k\}$ correspond to topics, from which we can derive probabilistic distributions of words $w \in W = \{w_1, ..., w_m\}$ in a document $d \in D = \{d_1, ..., d_n\}$ with $k << n$. Thus, the goal is to learn $P(z \mid d)$ and $P(w|z)$ by maximizing the log-likelihood function:

$$L(W, D) = \sum_{d \in D} \sum_{w \in W} n(w, d) \log P(w \mid d), \quad (2)$$

where $n(w, d)$ is the term frequency of w in d. Using Bayes' formula, the conditional probability $P(w \mid d)$ is defined as:

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d). \quad (3)$$

Using the Expectation Maximization (EM) algorithm (Dempster et al., 1977), we estimate the parameters $P(z|d)$ and $P(w|z)$ via an iterative process that alternates two steps: (i) an expectation step (E) in which posterior probabilities are computed for each latent topic z; and (ii) a maximization (M) step, in which the parameters are updated for the posterior probabilities computed in the previous E-step. Details of how to efficiently implement the re-estimation formulas can be found in Federico (2002).

Iterating the E- and M-steps will lead to a convergence that approximates the maximum likelihood equation in (2).

A document-topic distribution $\hat{\theta}$ can be inferred on a new document d' by maximizing the following equation:

$$\hat{\theta} = \arg\max_{\theta} \sum_{w} n(w, d') \log \sum_{z} P(w \mid z) \theta_{z,d'}, \quad (4)$$

where $\theta_{z,d'} = P(z \mid d')$. (4) can be maximized by performing Expectation Maximization on document d' by keeping fixed the word-topic distributions already estimated on the training data. Consequently, a word-document distribution can be inferred by applying the mixture model (3) (see Federico, 2002 for details).

4 MDI Adaptation

An n-gram language model approximates the probability of a sequence of words in a text $W_1^T = w_1, ..., w_T$ drawn from a vocabulary V by the following equation:

$$P(W_1^T) = \prod_{i=1}^{T} P(w_i|h_i), \quad (5)$$

where $h_i = w_{i-n+1}, ..., w_{i-1}$ is the history of $n - 1$ words preceding w_i. Given a training corpus B, we can compute the probability of a n-gram from a smoothed model via interpolation as:

$$P_B(w|h) = f_B^*(w|h) + \lambda_B(h) P_B(w|h'), \quad (6)$$

where $f_B^*(w|h)$ is the discounted frequency of sequence hw, h' is the lower order history, where $|h| - 1 = |h'|$, and $\lambda_B(h)$ is the zero-frequency probability of h, defined as:

$$\lambda_B(h) = 1.0 - \sum_{w \in V} f_B^*(w|h).$$

Federico (1999) has shown that MDI Adaptation is useful to adapt a background language model with a small adaptation text sample A, by assuming to have only sufficient statistics on unigrams. Thus, we can reliably estimate $\hat{P}_A(w)$ constraints on the marginal distribution of an adapted language model $P_A(h, w)$ which minimizes the Kullback-Leibler distance from B, i.e.:

$$P_A(\cdot) = \arg\min_{Q(\cdot)} \sum_{hw \in V^n} Q(h, w) \log \frac{Q(h, w)}{P_B(h, w)}. \quad (7)$$

The joint distribution in (7) can be computed using Generalized Iterative Scaling (Darroch and Ratcliff, 1972). Under the unigram constraints, the GIS algorithm reduces to the closed form:

$$P_A(h,w) = P_B(h,w)\alpha(w), \tag{8}$$

where

$$\alpha(w) = \frac{\hat{P}_A(w)}{P_B(w)}. \tag{9}$$

In order to estimate the conditional distribution of the adapted LM, we rewrite (8) and simplify the equation to:

$$P_A(w|h) = \frac{P_B(w|h)\alpha(w)}{\sum_{\hat{w}\in V} P_B(\hat{w}|h)\alpha(\hat{w})}. \tag{10}$$

The adaptation model can be improved by smoothing the scaling factor in (9) by an exponential term γ (Kneser et al., 1997):

$$\alpha(w) = \left(\frac{\hat{P}_A(w)}{P_B(w)}\right)^{\gamma}, \tag{11}$$

where $0 < \gamma \leq 1$. Empirically, γ values less than one decrease the effect of the adaptation ratio to reduce the bias.

As outlined in Federico (2002), the adapted language model can also be written in an interpolation form:

$$f_A^*(w|h) = \frac{f_B^*(w|h)\alpha(w)}{z(h)}, \tag{12}$$

$$\lambda_A(h) = \frac{\lambda_B(h)z(h')}{z(h)}, \tag{13}$$

$$z(h) = (\sum_{w:N_B(h,w)>0} f_B^*(w|h)\alpha(w)) + \lambda_B(h)z(h'), \tag{14}$$

which permits to efficiently compute the normalization term for high order n-grams recursively and by just summing over observed n-grams. The recursion ends with the following initial values for the empty history ϵ:

$$z(\epsilon) = \sum_w P_B(w)\alpha(w), \tag{15}$$

$$P_A(w|\epsilon) = P_B(w)\alpha(w)z(\epsilon)^{-1}. \tag{16}$$

MDI adaptation is one of the adaptation methods provided by the IRSTLM toolkit and was applied as explained in the following section.

5 Bilingual Latent Semantic Models

Similar to the treatment of documents in HM-BiTAM (Zhao and Xing, 2008), we combine parallel texts into a document-pair $(\mathbf{E},\mathbf{F})$ containing n parallel sentence pairs $(e_i, f_i), 1 < i \leq n$, corresponding to the source and target languages, respectively. Based on the assumption that the topics in a parallel text share the same semantic meanings across languages, the topics are sampled from the same topic-document distribution. We make the additional assumption that stop-words and punctuation, although having high word frequencies in documents, will generally have a uniform topic distribution across documents; therefore, it is not necessary to remove them prior to model training, as they will not adversely affect the overall topic distribution in each document. In order to ensure the uniqueness between word tokens between languages, we annotate **E** with special characters. We perform PLSA training, as described in Section 3.1 and receive word-topic distributions $P(w|z), w \in V_E \cup V_F$

Given an untranslated text $\hat{\mathbf{E}}$, we split $\hat{\mathbf{E}}$ into a sequence of documents D. For each document $d_i \in D$, we infer a full word-document distribution by learning $\hat{\theta}$ via (4). Via (3), we can generate the full word-document distribution $P(w \mid d)$ for $w \in V_F$.

We then convert the word-document probabilities into pseudo-counts via a scaling function:

$$n(w \mid d) = \frac{P(w \mid d)}{\max_{w'} P(w' \mid d)} \cdot \Delta, \tag{17}$$

where Δ is a scaling factor to raise the probability ratios above 1. Since our goal is to generate a unigram language model on the target language for adaptation, we remove the source words generated in (17) prior to building the language model.

From our newly generated unigram language model, we perform MDI adaptation on the background LM to yield an adapted LM for translating the source document used for the PLSA inference step.

6 Experiments

Our experiments were done using the TED Talks collection, used in the IWSLT 2010 evaluation task[1].

[1] http://iwslt2010.fbk.eu/

In IWSLT 2010, the challenge was to translate talks from the TED website[2] from English to French. The talks include a variety of topics, including photography and pyschology and thus do not adhere to a single genre. All talks were given in English and were manually transcribed and translated into French. The TED training data consists of 329 parallel talk transcripts with approximately 84k sentences. The TED test data consists of transcriptions created via 1-best ASR outputs from the KIT Quaero Evaluation System. It consists of 758 sentences and 27,432 and 27,307 English and French words, respectively. The TED talk data is segmented at the clause level, rather than at the level of sentences.

Our SMT systems are built upon the Moses open-source SMT toolkit (Koehn et al., 2007)[3]. The translation and lexicalized reordering models have been trained on parallel data. One 5-gram background LM was constructed from the French side of the TED training data (740k words), smoothed with the improved Kneser-Ney technique (Chen and Goodman, 1999) and computed with the IRSTLM toolkit (Federico et al., 2008). The weights of the log-linear interpolation model were optimized via minimum error rate training (MERT) (Och, 2003) on the TED development set, using 200 best translations at each tuning iteration.

This paper investigates the effects of language model adaptation via bilingual latent semantic modeling on the TED background LM against a baseline model that uses only the TED LM.

6.1 Bilingual Latent Semantic Model

Using the technique outlined in Section 5, we construct bilingual documents by splitting the parallel TED training corpus into 41,847 documents of 5 lines each. While each individual TED lecture could be used as a document, our experimental goal is to simulate near-time translation of speeches; thus, we prefer to construct small documents to simulate topic modeling on a spoken language scenario in which the length of a talk is not known a priori. We annotate the English source text for removal after inference. Figure 1 contains a sample document constructed for PLSA training. (In fact, we distinguish English words from French words by attaching to the former a special suffix.) By using our inhouse implementation, training of the PLSA model on the bilingual collection converged after 20 EM iterations.

robert lang is a pioneer of the newest kind of origami – using math and engineering principles to fold mind-blowingly intricate designs that are beautiful and , sometimes , very useful . my talk is " flapping birds and space telescopes . " and you would think that should have nothing to do with one another , but i hope by the end of these 18 minutes , you 'll see a little bit of a relation . robert lang est un pionnier des nouvelles techniques d' origami - basées sur des principes mathématiques et d' ingénierie permettant de créer des modèles complexes et époustouflants , qui sont beaux et parfois , très utiles . ma conférence s' intitule " oiseaux en papier et télescopes spatiaux " . et vous pensez probablement que les uns et les autres n' ont rien en commun , mais j' espère qu' à l' issue de ces 18 minutes , vous comprendrez ce qui les relie .

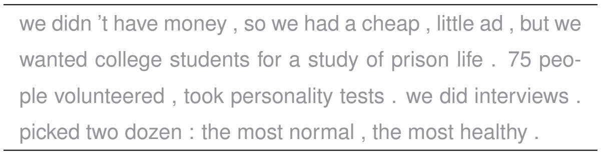
Figure 1: A sample bilingual document used for PLSA training.

Using our PLSA model, we run inference on each of the 476 test documents from the TED lectures, constructed by splitting the test set into 5-line documents. Since our goal is to translate and evaluate the test set, we construct monolingual (English) documents. Figure 2 provides an example of a document to be inferred. We collect the bilingual unigram pseudocounts after 10 iterations of inference and remove the English words. The TED lecture data is transcribed by clauses, rather than full sentences, so we do not add sentence splitting tags before training our unigram language models.

As a result of PLSA inference, the probabilities of target words increase with respect to the background language model. Table 1 demonstrates this phenomenon by outlining several of the top ranked words that have similar semantic meaning to non-stop words on the source side. In every case, the probability $P_A(w)$ increases fairly substantially with respect to the $P_B(w)$. As a result, we expect that the adapted language model will favor both fluent and semantically correct translations as the adaptation is suggesting better lexical choices of words.

[2] http://www.ted.com/talks/

[3] http://www.statmt.org/moses/



we didn 't have money , so we had a cheap , little ad , but we wanted college students for a study of prison life . 75 people volunteered , took personality tests . we did interviews . picked two dozen : the most normal , the most healthy .

Figure 2: A sample English-only document (#230) used for PLSA inference. A full unigram word distribution will be inferred for both English and French.

Rank	Word	$P_A(w)$	$P_B(w)$	$P_A(w)/P_B(w)$
20	gens	8.41E-03	4.55E-05	184.84
22	vie	8.30E-03	1.09E-04	76.15
51	prix	2.59E-03	8.70E-05	29.77
80	école	1.70E-03	6.13E-05	27.73
83	argent	1.60E-03	3.96E-05	40.04
86	personnes	1.52E-03	2.75E-04	5.23
94	aide	1.27E-03	7.71E-05	16.47
98	étudiants	1.20E-03	7.12E-05	16.85
119	marché	9.22E-04	9.10E-05	10.13
133	étude	7.63E-04	4.55E-05	16.77
173	éducation	5.04E-04	2.97E-05	16.97
315	prison	2.65E-04	1.98E-05	13.38
323	université	2.60E-04	2.97E-05	8.75

Table 1: Sample unigram probabilities of the adaptation model for document #230, compared to the baseline unigram probabilities. The French words selected are semantically related to the English words in the adapted document. The PLSA adaptation infers higher unigram probabilities for words with latent topics related to the source document.

6.2 MDI Adaptation

We perform MDI adaptation with each of the unigram language models to update the background TED language model. We configure the adaptation rate parameter γ to 0.3, as recommended in Federico (2002). The baseline LM is replaced with each adapted LM, corresponding to the document to be translated. We then calculate the mean perplexity of the adapted LMs and the baseline, respectively. The perplexity scores are shown in Table 2. We observe a 15.3% relative improvement in perplexity score over the baseline.

6.3 Results

We perform MT experiments on the IWSLT 2010 evaluation set to compare the baseline and adapted LMs. In the evaluation, we notice a 0.85 improvement in BLEU (%), yielding a 3% improvement over the baseline. The same performance trend in NIST is observed with a 2.4% relative improvement compared to the unadapted baseline. Our PLSA and MDI-based adaptation method not only improves fluency but also improves adequacy: the topic-based adaptation approach is attempting to suggest more appropriate words based on increased unigram probabilities than that of the baseline LM. Table 3 demonstrates a large improvement in unigram selection for the adapted TED model in terms of the individual contribution to the NIST score, with diminishing effects on larger n-grams. The majority of the overall improvements are on individual word selection.

Examples of improved fluency and adequacy are shown in Figure 3. Line 285 shows an example of a translation that doesn't provide much of an n-gram improvement, but demonstrates more fluent output, due to the deletion of the first comma and the movement of the second comma to the end of the clause. While "installation" remains an inadequate noun in this clause, the adapted model reorders the root words "rehab" and "installation" (in comparison with the baseline) and improves the grammaticality of the sentence; however, the number does not match between the determiner and the noun phrase. Line 597 demonstrates a perfect phrase translation with respect to the reference translation using semantic paraphrasing. The baseline phrase "d'origine" is transformed and attributed to the noun. Instead of translating "original" as a phrase for "home", the adapted model captures the original meaning of the word in the translation. Line 752 demonstrates an improvement in adequacy through the replacement of the word "quelque" with "autre." Additionally, extra words are removed.

These lexical changes result in the improvement in translation quality due to topic-based adaptation via PLSA.

LM	Perplexity	BLEU (%)	NIST
Adapt TED	162.44	**28.49**	**6.5956**
Base TED	191.76	27.64	6.4405

Table 2: Perplexity, BLEU, and NIST scores for the baseline and adapted models. The perplexity scores are averaged across each document-specific LM adaptation.

NIST	1-gram	2-gram	3-gram
Adapt TED	4.8077	1.3925	0.3229
Base TED	4.6980	1.3527	0.3173
Difference	0.1097	0.0398	0.0056

Table 3: Individual unigram NIST scores for n-grams 1-3 of the baseline and adapted models. The improvement of the adapted model over the baseline is listed below.

(Line 285)
, j' ai eu la chance de travailler dans les *installations , rehab*
j' ai eu la chance de travailler dans les *rehab installation* ,
j' ai la chance de travailler dans un centre de désintoxication ,
(Line 597)
d' origine , les idées qui ont de la valeur –
d' avoir des idées originales qui ont de la valeur –
d' avoir des idées originales qui ont de la valeur –
(Line 752)
un nom qui appartient à *quelque* chose *d' autre* , le soleil .
un nom qui appartient à *autre* chose , le soleil .
le nom d' une *autre* chose , le soleil .

Figure 3: Three examples of improvement in MT results: the first sentence in each collection corresponds to the baseline, the second utilizes the adapted TED LMs, and the third is the reference translation.

7 Conclusions

An alternative approach to bilingual topic modeling has been presented that integrates the PLSA framework with MDI adaptation that can effectively adapt a background language model when given a document in the source language. Rather than training two topic models and enforcing a one-to-one correspondence for translation, we use the assumption that parallel texts refer to the same topics and have a very similar topic distribution. Preliminary experiments show a reduction in perplexity and an overall improvement in BLEU and NIST scores on speech translation. We also note that, unlike previous works involving topic modeling, we did not remove stop words and punctuation, but rather assumed that these features would have a relatively uniform topic distribution.

One downside to the MDI adaptation approach is that the computation of the normalization term $z(h)$ is expensive and potentially prohibitive during continuous speech translation tasks. Further investigation is needed to determine if there is a suitable approximation that avoids computing probabilities across all n-grams.

Acknowledgments

This work was supported by the T4ME network of excellence (IST-249119), funded by the DG INFSO of the European Commission through the Seventh Framework Programme. The first author received a grant under the Erasmus Mundus Language & Communication Technologies programme.

References

David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 4(13): 359–393, 1999.

J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

Marcello Federico. Efficient language model adaptation through MDI estimation. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 4, pages 1583–1586, Budapest, Hungary, 1999.

Marcello Federico. Language Model Adaptation through Topic Decomposition and MDI Estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 703–706, Orlando, FL, 2002.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, pages 1618–1621, Melbourne, Australia, 2008.

George Foster and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0217`.

Zhengxian Gong, Yu Zhang, and Guodong Zhou. Statistical Machine Translation based on LDA. In *Universal Communication Symposium (IUCS), 2010 4th International*, pages 286 –290, oct. 2010. doi: 10.1109/IUCS.2010.5666182.

Thomas Hofmann. Probabilistic Latent Semantic Analysis. In *Proceedings of the 15th Conference on Uncertainty in AI*, pages 289–296, Stockholm, Sweden, 1999.

Bo-June (Paul) Hsu and James Glass. Style & topic language model adaptation using HMM-LDA. In *in Proc. ACL Conf. on Empirical Methods in Natural Language Processing – EMNLP*, pages 373–381, 2006.

Reinhard Kneser, Jochen Peters, and Dietrich Klakow. Language Model Adaptation Using Dynamic Marginals. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 1971–1974, Rhodes, Greece, 1997.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, 2007. URL `http://aclweb.org/anthology-new/P/P07/P07-2045.pdf`.

Philipp Koehn and Josh Schroeder. Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0233`.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, August 2009. URL `http://www.cs.umass.edu/~mimno/papers/mimno2009polylingual.pdf`.

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003. URL `http://www.aclweb.org/anthology/P03-1021.pdf`.

Abhinav Sethy, Panayiotis Georgiou, and Shrikanth Narayanan. Selecting relevant text subsets from web-data for building topic specific language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 145–148, New York City, USA, June 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N/N06/N06-2037`.

Yik-Cheung Tam and Tanja Schultz. Incorporating monolingual corpora into bilingual latent semantic analysis for crosslingual lm adaptation. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4821 –4824, april 2009. doi: 10.1109/ICASSP.2009.4960710.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21:187–207, December 2007. ISSN 0922-6567. doi: 10.1007/s10590-008-9045-2. URL `http://portal.acm.org/citation.cfm?id=1466799.1466803`.

Bing Zhao and Eric P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1689–1696. MIT Press, Cambridge, MA, 2008.

Bing Zhao, Matthias Eck, and Stephan Vogel. Language Model Adaptation for Statistical Machine Translation via Structured Query Models. In *Proceedings of Coling 2004*, pages 411–417, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.