Learning Latent Topic Information for Language Model Adaptation

Shixiang Lu*,**, Wei Wei, Xiaoyin Fu, Lichun Fan, and Bo Xu

Interactive Digital Media Technology Research Center
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Haidian District, Beijing 100190, China
shixiang.lu@ia.ac.cn

Abstract. This paper is concerned with data selection for adapting language model (LM) in statistical machine translation (SMT), and aims to find the LM training sentences that are topic similar to the translation task. Although the traditional methods have gained significant performance, they ignore the topic information and the distribution of words in calculating the sentence similarity. In this paper, the authors propose a topic model to discover the latent topics in the content of sentences, and combine the latent topic based similarity with TF-IDF into a unified framework for data selection. Furthermore, the authors combine a cross-lingual projecting method with the topic model, which makes the data selection depend on the source input directly. Large-scale experimental results demonstrate that the proposed approach significantly outperforms the traditional approaches on both LM perplexity and SMT performance.

Keywords: topic information, cross-lingual projection, data selection, language model adaptation, statistical machine translation.

1 Introduction

Over the past few years, selecting training data which are similar to the translation task from the large corpus has become an important approach to improve the performance of language model (LM) in statistical machine translation (SMT) [1-5]. This would empirically provide more accurate lexical probabilities, and thus better match the translation task at hand[5].

The major challenge for data selection is how to measure the similarity between the queried sentence and the LM training corpus. To solve this problem, many researchers proposed various kinds of similarity measures to select similar sentences for LM adaptation, such as TF-IDF[1-3, 6], centroid similarity[4], cross-entropy difference[5], cross-lingual information retrieval[7], and cross-lingual similarity (CLS)[8]. Unfortunately, they all take the similarity measure without considering the topic information and the distribution of words in the whole LM

* Contact author.

** This work was supported by 863 program in China (No. 2011AA01A207).

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 143–153, 2012.

training corpus. These information have been successfully used for LM adaptation in SMT[9, 10] and been proved very useful. This approach infers the topic posterior distribution of the source text, and then applies the inferred distribution to the target language LM via marginal adaptation. However, it focus on modify the LM itself, which is different from data selection method for LM adaptation.

To address this problem, we propose a more principled latent topic based data selection model for LM adaptation in SMT. To the best of our knowledge, this is the first extensive and empirical study of learning the latent topic information for data selection to adapt LM. We employ the topic model (e.g., Latent Dirichlet Allocation) to discover the latent topics in the whole content of LM training corpus. Then we calculate the topic-similarity between the first pass translation hypotheses[1] and the sentences in the LM training corpus based on the latent topic information. Moreover, we propose a cross-lingual projecting method, which projects the source input sentences in the translation task to the target language representation, and then we combine it with the topic model. Therefore, when given the source input sentence, we can select the topic-similar sentences directly without the first pass translation hypotheses. TF-IDF and latent topic information are based on different knowledge, we assume they are complementary to each other, and the performance can be further improved by combining them, as we will show in the experiments.

The remainder of this paper is organized as follows. The next section introduces some related work of LM adaptation. Section 2 describes our proposed latent topic based data selection model for LM adaptation. Section 3 presents large-scale experiments and analyses, and followed by conclusions and future work in section 4.

2 Related Work

A variety of latent topic models have been used for LM adaptation in speech recognition (SR)[11-19], which show the latent topic information are very useful for LM adaptation. The previous works have primarily focused on customizing a fixed n-gram LM for each lecture by combining n-gram statistics from general conversational speech, other lectures, textbooks, and other resources related to the target lecture[11-14]. Moreover, they focus on in-domain adaptation using large amounts of matched training data[19]. However, most, if not all, of the data available to train an LM in SMT are cross-topic and cross-style. Therefore, these previous latent topic based LM adapting methods in SR are not suitable for SMT, and we will illustrate a novel latent topic based data selection model for LM adaptation in this paper.

To the best of our knowledge, none of the existing studies have addressed data selection for LM adaptation in SMT by learning the latent topics. In the next

[1] Following [2, 4], we call the initial translations hypotheses which are generated by the baseline SMT system as the firs pass translation hypotheses.

section, we explore a new approach to discover the latent topic information into the similar data selection for LM adaptation.

3 Latent Topic Based Data Selection for LM Adaptation

For the first pass translation hypotheses or the source input sentences in the translation task, we estimate the bias LM, from the corresponding similar LM training sentences. Since this size of selected sentences is small, the corresponding bias LM is specific and more effective, giving high probabilities to those phrases that occur in the selected sentences.

The generic LM $P_g(w_i|h)$ and the bias LM $P_b(w_i|h)$ is combined using linear interpolation as adapted LM $P_a(w_i|h)$ [2,7], which is shown to improve performance over the individual models:

$$P_a(w_i|h) = \gamma P_g(w_i|h) + (1-\gamma)P_b(w_i|h) \tag{1}$$

where the interpolation factor γ can be simply estimated using the Powell Search algorithm[20] via cross-validation, and the bias LM is of the same order and smoothing algorithm as the generic LM.

The resulting adapted LM is then used in place of the generic LM in the translation process, would empirically provides more accurate lexical probabilities, and thus better matches the translation task at hand. Our work focuses on latent topic based data selection model, and the quality of this model is crucial to the performance of adapted LM.

3.1 Latent Topic Based Data Selection Model

Before introducing our proposed method, we first briefly describe the LDA model[21]. LDA models the generation of document content as two independent stochastic processes by introducing latent topic space. For an arbitrary word w in document d, (1) a topic z is first sampled from the multinomial distribution θ_d, which is generated from the Dirichlet prior parameterized by α; (2) and then the word w is generated from multinomial distribution ϕ_z, which is generated from the Dirichlet prior parameterized by β. The two Dirichlet priors for documents-topic distribution θ_d and topic-word distribution ϕ_z reduce the probability of overfitting training documents and enhance the ability of inferring topic distribution for new documents.

In latent topic based data selection model (LT), the first pass translation hypotheses and the sentences in the LM training corpus can be considered as documents. In this paper, we employ state-of-the-art topic model - LDA to discover the latent topics information and the distribution of words in them. We consider the first pass translation hypotheses as a question sentence s, and assume that s and the LM training sentence S are represented by a distribution over topics. $|s|$ represents the length of s, and we obtain the topic distribution of s by merging the topic distributions of words:

$$P_{LT}(z|s) = \frac{1}{|s|}\sum_{w \in s} P(z|w) \tag{2}$$

Then, we assume that s and S have the same prior probability, K represents the number of topics, N represents the numbers of s, so the score function can be written as:

$$\begin{aligned} P_{LT}(s|S) &= \sum_{z} P_{LT}(s|z)P_{LT}(z|S) \\ &= \sum_{z \in K} \frac{P_{LT}(z|s)P(s)}{P(z)} P_{LT}(z|S) \\ &= \frac{K}{N} \sum_{z \in K} P_{LT}(z|s)P_{LT}(z|S) \end{aligned} \tag{3}$$

3.2 Parameter Estimation

After introducing our proposed LT method, we will describe how to estimate the parameter used in the model. In LT, we introduce the new parameters, which lead to the inference not be done exactly. Expectation-Maximum (EM) algorithm is a possible choice for estimating the parameters of models with latent variables. However, EM suffers from the possibility of running into local maxima and the high computational burden. Therefore, we employ an alternative approach - Gibbs sampling[22], which is gaining popularity in recent work on latent topic analysis.

After training the model, we can get the following parameter estimations as:

$$\hat{\theta}_{sz} = \frac{n_{sz} + \alpha_z - 1}{\sum_{z'=1}^{K}(n_{sz'} + \alpha_{z'}) - 1} \tag{4}$$

$$\hat{\phi}_{zw} = \frac{n_{zw} + \beta_w - 1}{\sum_{v=1}^{V}(n_{zv} + \beta_v) - 1} \tag{5}$$

where n_{sz} and n_{zw} are the number of times of sentence s and word w which are assigned to the topic z, and V represents the number of unique words.

Next, we concentrate on how to select proper topic number to obtain our model with best performance and enough iteration to prevent the overfitting problem. We calculate the perplexity on LM training corpus C to estimate the performance of our model, which is a sequence of tuples $(s, w) \in C$:

$$Perplexity(C) = exp\{-\frac{\sum_{(s,w)\in C} lnP(w|s)}{|C|}\} \tag{6}$$

where, the probability $P(w|s)$ is calculated as follow:

$$P(w|s) = \sum_{z=1}^{K} P(w|z)P(z|s) \tag{7}$$

3.3 Combining Latent Topic with TF-IDF for Data Selection

Since the LT model and TF-IDF use different strategies for data selection, we assume that this two models are complementary to each other, it is interesting to explore how to combine their strength. In this section, we propose an approach to linearly combine the LT model with the TF-IDF model for data selection. In this paper, we choose TF-IDF as the foundation of our solution since TF-IDF has gained significant performance for LM adaptation in SMT[1-3, 6]. Formally, we have

$$P_{LT\_TF-IDF}(s|S) = \mu P_{LT}(s|S) + (1-\mu)P_{TF-IDF}(s|S) \tag{8}$$

where, the relative importance of LT and TF-IDF is adjusted through the interpolation parameter μ.

3.4 Latent Topic Based Cross-Lingual Data Selection Model

Inspired by the work of CLS[8], we assume the following processing. The source sentence u and the target sentence v lie in two different vector space, we need to find a projection of u in the target vocabulary vector space before similarity can be evaluated. We estimate the bilingual word co-occurrence matrix Σ from an unsupervised, automatic word alignment induced over the SMT parallel training corpus. We use the GIZA++ toolkit to estimate the parameters of IBM Model 4, and combine the forward and backward viterbi alignments. Then, the projection of the source sentence u in the target vector space can be calculated by the vector-matrix product, as show:

$$\hat{v} = u\Sigma \tag{9}$$

The target term in $\hat{v}$ will be emphasized that most frequently co-occur with the source term in u. $\hat{v}$ can be interpreted as a "bag of words" translation of u. Next, we extend $\hat{v}$ into latent topic based cross-lingual data selection model (CLLT) for LM adaptation. We consider $\hat{v}$ as the first pass translation hypotheses $\hat{s}$, so CLLT can be written as follows:

$$P_{CLLT}(\hat{s}|S) = \frac{K}{N}\sum_{z\in K} P_{CLLT}(z|\hat{s})P_{CLLT}(z|S) \tag{10}$$

We use CLS to calculate the source sentence u to each target sentence S. However, due to the lack of optimization measures for sparse vector representation, the similarity is not accurate. In our model, we add the optimization measures (TF-IDF), called CLS_s, which improves the performance, as we will show in the experiment. What is more, we apply this criterion for the first time to the task of cross-lingual data selection for LM adaptation in SMT. This model can be written as follow:

$$\begin{aligned} P_{CLS_s}(\hat{s}|S) &= \frac{S^T \cdot \hat{s}}{\|S\|\|\hat{s}\|} \\ &= \frac{S^T \cdot u\Sigma}{\|S\|\|u\Sigma\|} \end{aligned} \tag{11}$$

Lastly, we combine CLLT and CLSs into a cross-lingual data selection framework by the linear interpolation parameter, as follows:

$$P_{CLLT\_CLS_s}(\hat{s}|S) = \lambda P_{CLLT}(\hat{s}|S) + (1-\lambda)P_{CLS_s}(\hat{s}|S) \quad (12)$$

where, the relative importance of CLLT and CLS_s is adjusted through the interpolation parameter λ.

4 Experiments and Results

We measure the utility of the proposed LM adaptation approach and the traditional approaches in two ways: (a) comparing the reference translations based perplexity of adapted LMs with the generic LM, and (b) comparing SMT performance of adapted LMs with the generic LM.

4.1 Corpus

We conduct experiments on two Chinese-to-English translation tasks: IWSLT-07 (dialogue domain) and NIST-06 (news domain).

IWSLT-07. The bilingual corpus comes from BTEC and CJK corpus, which contains 3.82K sentence pairs. The LM training corpus is from the English side of the parallel data (BTEC, CJK and CWMT2008), which consists of 1.34M sentences. IWSLT-07 test set consists of 489 sentences with 4 English reference translations each, and development set is the IWSLT-05 test set with 506 sentences.

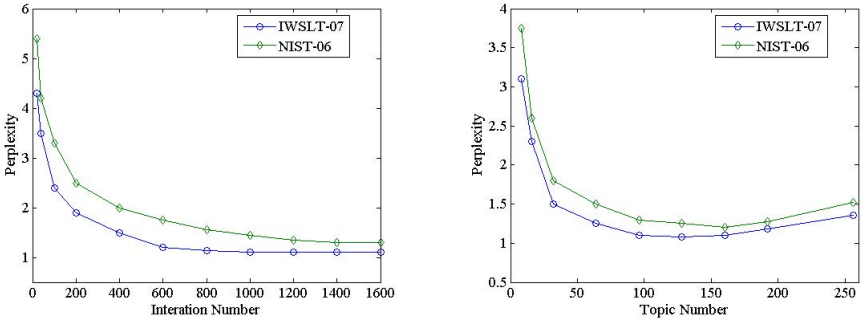
NIST-06. The bilingual corpus comes from LDC[2], which consists of 3.4M sentence pairs. The LM training corpus is from the English side of the English Gigaword corpus[3], which consists of 11.3M sentences. NIST-06 MT Evaluation test set consists of 1664 sentences with 4 English reference translations each, and development set is NIST-05 MT Evaluation test set with 1084 sentences.

4.2 Iteration and Topic Number Selection

Fig. 1(a) shows the influence of iteration number of Gibbs sampling on the topic model generalization ability. Empirically, we set the topic number as 96 on IWSLT-07 and 168 on NIST-06, respectively, and change the iteration number in the experiments. Note that the lower perplexity value indicates better generalization ability on the holdout LM training corpus. We see that the perplexity values decreases when the iteration times are below 1000 on IWSLT-06 and 1400 on NIST-06, respectively. Fig. 1(b) shows the perplexity values for different settings of the topic number. We see that the perplexity decreases when the number

[2] LDC2002E18, LDC2002T01, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07, LDC2004T08, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006T04, LDC2007T09

[3] LDC2007T07



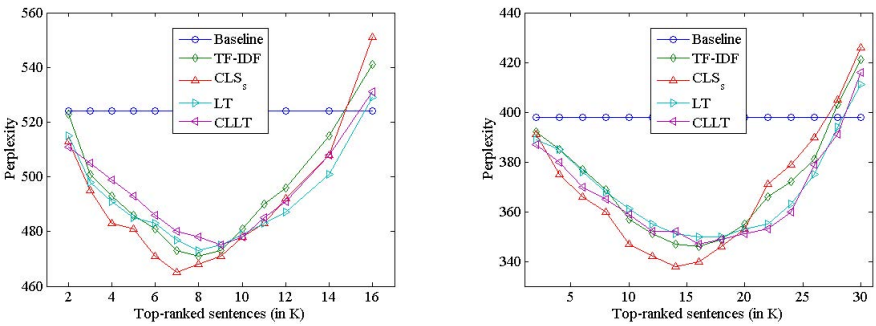
(a) interation number (b) topic number

Fig. 1. Perplexity vs. the number of different iterations and topics on two LM training corpus

of topics starts to increase. However, after a certain point, the perplexity values start to increase. Based on the above experiments, we train our latent topic model using (a) 96 topics and 1000 iterations on IWSLT-07 and (b) 168 topics and 1400 iterations on NIST-06, respectively.

4.3 Perplexity Analysis

We randomly divide the development set into five subsets and conduct 5-fold cross-validation experiments. In each trial, we tune the parameter γ in Equation (1) with four of five subsets and then apply it to one remaining subset. The experiments reported below are those averaged over the five trials.



(a) IWSLT-07 (b) NIST-06

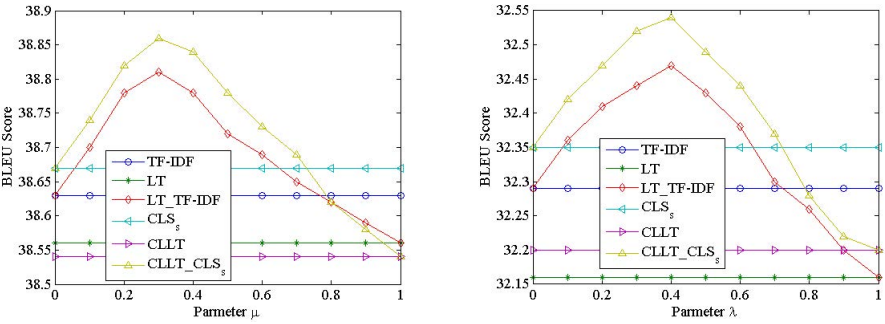
Fig. 2. English reference translation based perplexity of adapted LMs vs. the size of selected data on two test sets

For both IWSLT-07 and NIST-06, we estimate the generic 4-gram LM with the entire LM training corpus as a baseline. Then, we apply the proposed method and other traditional methods to select the top-N similar sentences which are similar to the test set, train the bias 4-gram LMs (with the same n-gram cutoffs tuned as above) with these selected sentences, and interpolate with the generic 4-gram LM as the adapted LMs. All the LMs are estimated using the SRILM toolkit[23]. Perplexity is a metric of LM performance, the lower values indicates the better performance. So we estimate the perplexity of English reference translation according to adapted LMs.

Fig. 2 shows the LM perplexity experiment results. We can see that the English reference translation based perplexity of adapted LMs decreases consistently when the size of selected top-N sentences increases, and also increases consistently after a certain size in all approaches. So proper size of similar sentences with the translation task make the LM perform well, but if too much noisy data take into the selected sentences, the performance become worse. Similar observations have been done by the previous work[1, 5]. The experiment results indicate that adapted LMs are significantly better predictors of the corresponding translation task at hand than the generic baseline LM.

4.4 Translation Experiments

To show the detailed performance of selected training data for LM adaptation in SMT, we carry out the later translation experiments with the lowest perplexity situation according to the above perplexity experiment, top 8K sentences on IWSLT-07 and top 16K sentences on NIST-06. We conduct translation experiments by HPB SMT[24] system, as to demonstrate the utility of LM adaptation in improving SMT performance by the BLEU[25] score, and use minimum error rate training[26] to tune the feature weights for maximum BLEU on the development set.



(a) IWSLT-07 (b) NIST-06

Fig. 3. The impact of parameters μ and λ to SMT performance on two development sets

Fig. 3 illustrates the impact results of parameters μ and λ to SMT performance on two development sets. TF-IDF, CLSs, LT and CLLT are used for reference. We see that the combined model LT_TF-IDF and CLLT_CLS$_s$ perform better than each other alone when they are between 0 and 0.6, the best performance gains when they are 0.3 on IWSLT-07 and 0.4 on NIST-06, and we use these parameters on two test sets.

Table 1. SMT performance with different data selection models for LM adaptation on two test sets

Method	#	BLEU	
		IWSLT-07	NIST-06
Baseline	1	33.60	29.15
TF-IDF	2	34.14	29.78
CLS	3	34.08	29.73
CLS$_s$	4	34.18	29.84
LT	5	34.07	29.65
CLLT	6	34.05	29.69
LT_TF-IDF	7	**34.32**	**29.96**
CLLT_CLS$_s$	8	**34.37**	**30.03**

Table 1 shows the main SMT performance of LM adaptation. The improvements are statistically significant at the 95% confidence interval, and we see some clear trends:

(1) Our proposed CLS$_s$ performs better than CLS (row 4 vs. row 3), because of the added smoothing measure which makes similarity computation more accurate.

(2) Our proposed LT and CLLT do not outperform the baseline method TF-IDF (row 5 and row 6 vs. row 2). This demonstrates that the knowledge extracted from LT is not as effective as that extracted from TF-IDF. However, LT models word-topic information and word-distribution information from the whole LM training corpus. The knowledge extracted from LT is much noisier than that of TF-IDF. We suspect the above reason leads to the poor performance of LT and CLLT.

(3) Our proposed LT_TF-IDF significantly outperforms LT and TF-IDF (row 7 vs. row 2 and row 5), and CLLT_CLSs significantly outperforms CLLT and CLSs (row 8 vs. row 4 and row 6). This demonstrates that the latent word-topic and word-distribution information extracted from LT is complementary to the knowledge extracted from TF-IDF on data selection for LM adaptation.

(4) Our proposed CLLT_CLSs outperforms LT_TF-IDF (row 8 vs. row 7), and CLSs outperforms TF-IDF (row 4 vs. row 2). This demonstrates that the first pass translation hypotheses have lots of noisy data[27, 28], which mislead the selected similar sentences[9, 16, 27, 28], and take noisy data into the selected sentences. However, cross-lingual data selection model can avoid this problem,

and makes the sentence selection depend on the source input translation task directly.

5 Conclusions and Future Work

In this paper, we propose a novel latent topic based data selection model for LM adaptation in SMT. Furthermore, we expand it into cross-lingual data selection by a cross-lingual projection. Compared to the traditional approaches, our approach is more effective because it takes the distribution of words and the latent topic information into the similarity measure. Large-scale experiments conducted on LM perplexity and SMT performance demonstrate that our approach significantly outperforms the traditional methods.

There are some extensions of this work in the future. We will utilize our approach to mine large-scale corpora by distribute infrastructure system, and investigate the use of our approach for other domains, such as speech translation systems.

References

1. Eck, M., Vogel, S., Waibel, A.: Language model adaptation for statistical machine translation based on information retrieval. In: Proceedings of LREC, pp. 327–330 (2004)
2. Zhao, B., Eck, M., Vogel, S.: Language model adaptation for statistical machine translation with structured query models. In: Proceedings of COLING, pp. 411–417 (2004)
3. Kim, W.: Language model adaptation for automatic speech recognition and statistical machine translation. Ph.D. thesis, The Johns Hopkins University (2005)
4. Masskey, S., Sethy, A.: Resampling auxiliary data for language model adaptation in machine translation for speech. In: Proceedings of ICASSP, pp. 4817–4820 (2010)
5. Axelrod, A., He, X., Gao, J.: Domain adaptation via pseudo in-domain data selection. In: Proceedings of EMNLP, pp. 355–362 (2011)
6. Foster, G., Kuhn, R.: Mixture-model adaptation for SMT. In: Proceedings of ACL, pp. 128–135 (2007)
7. Snover, M., Dorr, B., Marcu, R.: Language and translation model adaptation using comparable corpora. In: Proceedings of EMNLP, pp. 857–866 (2008)
8. Ananthakrishnan, S., Prasad, R., Natarajan, P.: On-line language model biasing for statistical machine translation. In: Proceedings of ACL, pp. 445–449 (2011)
9. Tam, Y.-C., Lane, I., Schultz, T.: Bilingual-LSA based LM adaptation for spoken language translation. In: Proceedings of ACL, pp. 520–527 (2007)
10. Tam, Y.-C., Lane, I., Schultz, T.: Bilingual-LSA based adaptation for statistical machine translation. Machine Translation 21, 187–207 (2008)
11. Nanjo, H., Kawahara, T.: Unsupervised language model adaptation for lecture speech recognition. In: Proceedings of ICSLP (2002)
12. Nanjo, H., Kawahara, T.: Language model and speaking rate adaptation for spontaneous presentation speech recognition. IEEE Tran. SAP 12(4), 301–400 (2004)
13. Leeuwis, E., Federico, M., Cettolo, M.: Language modeling and transcription of the TED corpus lectures. In: Proceedings of ICASSP (2003)

14. Park, A., Hazen, T., Glass, J.: Automatic processing of audio lectures for information retrieval: vocabulary selection and language modeling. In: Proceedings of ICASSP (2005)
15. Tam, Y.-C., Schultz, T.: Dynamic language model adaptation using variational bayes inference. In: Proceedings of INTEERSPEECH, pp. 5–8 (2005)
16. Tam, Y.-C., Schultz, T.: Unsupervised language model adaptation using latent semantic marginals. In: Proceedings of ICSLP, pp. 2206–2209 (2006)
17. Heidel, A., Chang, H.-A., Lee, L.-S.: Language model adaptation using latent dirichlet allocation and an efficient topic inference algorithm. In: Proceedings of INTERSPEECH (2007)
18. Chen, K.-Y., Chiu, H.-S., Chen, B.: Latent topic modeling of word vicinity information for speech recognition. In: Proceedings of ICASSP, pp. 5394–5397 (2010)
19. (Paul) Hsu, B.-J., Glass, J.: Style & topic language model adaptation using HMM-LDA. In: Proceedings of EMNLP, pp. 373–381 (2006)
20. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C. Cambridge Univ. Press (1992)
21. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
22. Gtiffiths, T.L.: Gibbs sampling in the generative model of latent dirichlet allocation (2002), `http://wwwpsych.stanford.edu/gruffydd/cogsci02/lda.ps`
23. Stolcke, A.: SRILM - An extensible language modeling toolkit. In: Proceedings of ICSLP, pp. 901–904 (2002)
24. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of ACL (2005)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of ACL, pp. 311–318 (2002)
26. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of ACL, pp. 160–167 (2003)
27. Wei, B., Pal, C.: Cross lingual adaptation: an experiment on sentiment classifications. In: Proceedings of ACL, pp. 258–262 (2010)
28. Lu, S., Wei, W., Fu, X., Xu, B.: Translation model based cross-lingual language model adaptation: from word models to phrase models. In: Proceedings of EMNLP-CoNLL, pp. 512–522 (2012)