Selection-Based Language Model for Domain Adaptation using Topic Modeling

Tsuyoshi Okita
School of Computer Science
Dublin City University
Glasnevin, Dublin 9, Ireland
tokita@computing.dcu.ie

Josef van Genabith
CNGL, School of Computer Science
Dublin City University
Glasnevin, Dublin 9, Ireland
josef@computing.dcu.ie

Abstract

This paper introduces a selection-based LM using topic modeling for the purpose of domain adaptation which is often required in Statistical Machine Translation. The performance of this selection-based LM slightly outperforms the state-of-the-art Moore-Lewis LM by 1.0% for EN-ES and 0.7% for ES-EN in terms of BLEU. The performance gain in terms of perplexity was 8% over the Moore-Lewis LM and 17% over the plain LM.

1 Domain Adaptation in Statistical Machine Translation

Domain adaptation is one important research area in Statistical Machine Translation (SMT) as well as other areas of NLP such as parsing. Domain adaptation tries to ensure that the performance is not radically decreased even if we translate a text in a test set whose genre is different from the parallel corpus which is used to build the system. Without loss of generality, the decoder of an SMT system can be written in the form of the noisy channel model $\arg\min_E P(E|F)P_{LM}(E)$ where the two components, the set of $P(E|F)$ and that of $P_{LM}(E)$, are the targets that we do *domain adaptation* on: the set of $P(E|F)$ is called a phrase table (or a rule table) and that of $P_{LM}(E)$ is called a language model (for simplicity, the model is written in the simplest form without indices). Hence, one approach to domain adaptation in SMT aims at obtaining a domain-adapted phrase table and language model [1]. In particular, it is observed in several papers that the domain adaptation of the language model is often the most effective route in domain adaptation. In this context, we explore domain-adapted language models using topic modeling [2] in this paper. Note that there is an alternative approach which applies transfer learning [3] for domain adaptation, which is not pursued in this paper. In the following, we focus on the domain adaptation of language models and we leave the topic of translation model domain adaptation as further work.[1]

The special setting for SMT would be the following: (1) there is a tendency that if a training corpus becomes big, e.g. more than a million sentences, we may need to think about the corpus as a combination of different genres, and (2) we may have some information about the genre of a test set as a whole or for each sentence (it is rare that we do not have any information about the genre of the test set).

2 Selection-Based Language Models

Let us prepare n kinds of language models $\{P_{LM_1}, \ldots, P_{LM_n}\}$ (we sometimes call this "a pool of language models" in the following) and a selection function $f(s)$ where s denotes a test sentence and

[1] The topic modeling for translation model can be found in [4]. Main differences are the usage of cross-entropy and interpolation. The topic modeling for system combination in SMT can be found in [5].

the discrete value $\{1, \ldots, n\}(= f(s))$ which is the outcome of selection function $f(s)$ indicates the indices of one of the language models $\{P_{LM_1}, \ldots, P_{LM_n}\}$. Before we describe the detailed method how to prepare such n kinds of language models using topic modeling (see the next section), we first describe the overall framework of selection-based language models.

First, the selection-based language model is processed in the following way. When a test sentence s_i arrives at the SMT decoder, by calculating a selection function $f(s_i)(= t)$ one LM P_{LM_t} is selected among the pool of language models $\{P_{LM_1}, \ldots, P_{LM_n}\}$. The selected LM is used in the SMT decoder only for the sentence s_i. When a different sentence s_{i+1} arrives, another LM may be selected by $f(s_{i+1})(= t)$, which can be different from the selection for the previous sentence P_{LM_t}. Since the standard SMT decoder uses only a single language model, the selection-based language model has much flexibility to choose for the particular sentence: one LM is selected during run time by the prepared selection function $f(s_i)$ from the n different language models.

Second, the performance measure of a selection-based language model should incorporate the fact that different LMs can be selected for each sentence. This is since the usual measure of perplexity of language models is measured irrelevant to the topic ID of each test sentence. Hence, the perplexity becomes much higher than the reality (if we fix picking up some LM to calculate the perplexity). In order to reflect that we switch the LM depending on the sentence, we need to update the definition of perplexity accordingly. In the domain adaptation context, the language models that we built have problems in its measure how to make a comparison in a convenient manner due to this nature which captures the characteristics in the domain adaptation context.

2.1 Selection-based Language Models by Topic Model

One example of selection-based language model can be built using topic models where we use Latent Dirichlet Allocation (LDA) [2]. This paper explores this method. Since LDA is a representative model in the larger category of LDA including the structured LDA [6] and the correlated LDA [7], this method is principally applicable to any topic models.

Suppose we fix k and apply topic modeling on the data which consists of training and development corpora.[2] The first step is to prepare the k (sentence-based) clustered corpora. Using LDA which represents topics as multinomial distributions over the k unique word-types in the training and development corpora and represents documents as a mixture of topics, we obtain k separated topic words on each word (from now on, we call them the topic ID 1 to the topic ID k). For a given number of topics k and number of sentences N, we assume that a decent number of (sentence-based) split of corpus would be k if $k < N$, otherwise N.[3] Considering the topic distribution in each sentence under this assumption where the same topic ID suggests the closer topic distributions, we obtain k separate sentences for training and development corpora. Hence, we do indexing of all the data, i.e. training and development corpora, with the label of topic ID i ($1 \leq i \leq k$). We refer to each LM with the topic ID from 1 to k in the following. The second step is to prepare the general-domain corpus. This corpus is necessary to calculate the cross entropy (hence, similarity and dis-similarity is considered), which is the state-of-the-art domain adaptation technique [1, 8]. One approach to select such a general domain corpus from the training corpus would be to check the constituency of each sentence where the high frequency words make up more than 60% of the words in the sentence and subsample the general domain corpus from this. We set the threshold 60% in order that we avoid the situation where the sentence consists of mostly functional words if the threshold is too high.

The third step is to obtain a pool of language models by the prepared k separate corpora and the general domain corpus. Although one approach would be simply to use k separate sentences only for building each language model P_{LM_k}, this method runs into coverage problems, i.e. it is possible that we come across a lot of Out-of-Vocabulary words (OOV words) in the unseen test set, if the original corpus is not big enough. We take an alternative approach based on LM interpolation [9] with all other LMs keeping the weight of a particular language model P_{LM_i} big while others are small. We do this interpolation tuned on the development corpus of topic ID i and the general domain corpus by taking the cross entropy [1, 8] and we obtain the weights for each language model

[2] In the strict setting, we will first train the model of topic modeling by training and development corpora. Then, we infer the topic ID by the trained model for the test corpora.

[3] Note that we did not do the experiment of latter category of assumption since k was at most 50 in our experiment while N was around 3000.

$\sum w_i P_{LM_i} (= P_{LM_{i'}})$ and obtain the interpolated LM i'. Note that when we do the tuning using the development corpus of topic ID i, we call the resulting language model the topic ID i-adapted language model. Note also that a dash $'$ denotes the *interpolated* LM. We do this interpolation for each topic ID from 1 to k and we obtain the pool of interpolated LMs $\{P_{LM_{1'}}, \ldots, P_{LM_{k'}}\}$. The overall procedure is as follows.

2.2 Perplexity for Selection-based Language Model

The definition of perplexity by SRILM toolkit [9] [4] is shown as in (1) where we have N sentence in a test set:

$$\text{perplexity} = 10 * *(\sum_{i=1}^{N} \log P_{LM}(s_i)/[\sum_{i=1}^{N}(\text{len}(s_i) - \text{oov}(s_i)) + N]) \quad (2)$$

where $\log P_{LM}(s_i)$ denotes the log probability of sentence s_i, len(s_i) denotes the total words in sentence s_i, and oov(s_i) denotes the number of OOVs in sentence s_i. In order to incorporate the definition of the selection-based language model, the calculation of the log probability logProb needs to be done for each *domain adapted* language model i. Let $\{s_1, \ldots, s_m\}$ denote a test set of m sentences and $f(s)$ denote a selection function (or the inference function using the trained topic model). Hence, this can be rewritten as in (3): [5]

$$\text{perplexity} = 10 * *(\sum_{i=1}^{N} \log P_{LM_{f(s_i)}}(s_i)/[\sum_{i=1}^{N}(\text{len}(s_i) - \text{oov}(s_i)) + N]) \quad (3)$$

where s_i denotes the i-th sentence ($0 \leq i \leq N$). This definition can be read in the following way: depending on the i-th sentence s_i, the topic ID is selected by $f(s_i)$, hence the corresponding language model $P_{LM_{f(s_i)}}$. Using this language model $P_{LM_{f(s_i)}}$, the log probability of all the words in a sentence s_i is calculated. Then, such log probabilities from all the sentences are summed.

3 Experimental Results

Intrinsic Evaluation The first experiment is an intrinsic evaluation of selection-based language models. We evaluate this with the perplexity which is described in Section 2.2. The experimental setting is as follows. We use a set of randomly sampled 200k sentences in the English side of the ES-EN Europarl.[6] We expect that this genre can be devided into several smaller genres in news domain although this corpus is not a concatenation of several genres.[7] We use LDA of mallet [10] where we set the parameter of the Dirichlet priors $\alpha = 0.01$ and $\beta = 0.01$.[8] We use SRILM [9] for LM building as well as the LM interpolation with modified Kneser-Ney smoothing with / without pruning.

[4] This definition is written on the FAQ page of SRILM as in (1):

$$\text{perplexity} = 10 * *(\sum_{i=1}^{N} \text{logProb}_i/[\sum_{i=1}^{N}(\text{numWords}_i - \text{numOOV}_i) + \text{numSent}]) \quad (1)$$

where we slightly change the notation.

[5] Suppose that we have a testset consisting of two sentences. The sentence-based perplexity of the first sentence (1 sentences, 14 words, 0 OOVs, 0 zeroprobs, logprob= -47.7842 ppl= 1533.26 ppl1= 2589.16) using LM1 was calculated by $10 * *(47.7842/(14 - 0 + 1))$. The same figures for the second sentence (1 sentences, 6 words, 0 OOVs, 0 zeroprobs, logprob= -21.1041 ppl= 1034.84 ppl1= 3291.17) but using LM2 is $10 * *(21.1041/(6 - 0 + 1))$. Then, the overall perplexity which we use the first sentence by LM1 and the second sentence by LM2 is calculated by $10 * *((47.7842 + 21.1041)/(14 + 6 + 1 + 1)) = 1352.97$. Hence, this will become (2 sentences, 20 words, 0 OOVs, 0 zeroprobs, logprob= -68.8883 ppl= 1352.97 ppl1= 2782.38).

[6] http://www.statmt.org

[7] The experiment with the corpus which is a concatenation of several genres of corpora will be a further work.

[8] α is a prior on the topic distributions for document i, and β is a prior on the word distribution for topic k [2].

Figure 1 shows the results. The performance of Selection-based LMs (light blue and dark blue) are better than Moore-Lewis LM and plain LM. Especially, the performance with the topic cluster number of k=3, k=12, k=19 achieved the best performance of 8% (relative) improvement over Moore-Lewis LM and 17% improvement over the plain LM.

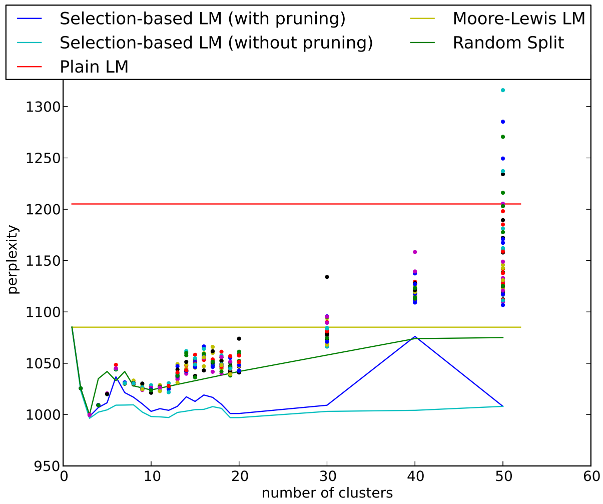

Figure 1: Figure shows the perplexity comparison of various methods. Dots represent topic clusters. Plain LM and Moore-Lewis LM are constant since they are not affected by the number of topic clusters. Two selection-based LM and random split are affected by the number of topic clusters.

Extrinsic Evaluation The second experiment is extrinsic evaluation using the SMT systems. There is one additional procedure for this experiment. Preparing the parallel corpus, i.e. ES-EN Europarl,[9] we concatenate the corresponding English and Spanish sentence pairs in a sentence and perform LDA. By this procedure we obtain a selection function $g(s)(= t)$ on the source side. Since only the source side sentences of the test set are shown to the SMT decoder, we place this function in the position of $f(s)$.

We build SMT system using Moses [11]. We use SRILM [9] for LM building as well as the LM interpolation with the modified Kneser-Ney smoothing with / without pruning. Our experiment is evaluated by BLEU [12]. We use the resources available for ES-EN language pair at the WMT13 site:[10] parallel corpora consist of Europarl [11] with 1,966k sentence pairs, UN corpus with 11,196k sentence pairs, news commentary corpus with 174k sentence pairs, and common crawl corpus with 1,845k sentence pairs, and monolingual corpora for English consist of Europarl [11] with 2,218k sentences and news language model data with 13,384k sentences.

Table 1 shows the results where the rows for 'SelectionLM' show the results of our method. In both directions EN-ES and ES-EN, the relative BLEU improvement between SelectionLM and plainLM was 2.4% in EN-ES and 2.0% in ES-EN, the relative improvement between SelectionLM and Moore-Lewis LM was 1.0% in EN-ES and 0.7% in ES-EN. Note that we choose k heuristically: k=3, k=4, k=5 were tried. The table shows that the results for NIST follows that for BLEU: the relative NIST improvement between SelectionLM and plainLM was 2.0% in EN-ES and 1.9% in ES-EN, while that between SelectionLM and Moore-Lewis LM was 0.9% in EN-ES and 0.5% in ES-EN. This result is statistically significant by the paired bootstrap resampling [13].

4 Conclusion

This paper introduces a selection-based LM using topic modeling. The performance of this selection-based LM slightly outperforms the state-of-the-art Moore-Lewis LM by 1.0% for EN-

[9] http://www.statmt.org

[10] http://www.statmt.org/wmt13.

	PBSMT EN-ES			PBSMT ES-EN		
	SelectionLM	Moore-Lewis	plain LM	SelectionLM	Moore-Lewis	plain LM
topic model	Y	N	N	Y	N	N
BLEU	29.7	29.4	29.0	30.5	30.3	29.9
NIST	7.70	7.63	7.52	7.73	7.69	7.58

Table 1: Table shows the score using three kinds of LM: SelectionLM which is our LM, the Moore-Lewis LM (state-of-the-art), and the plain LM. SelectionLM uses the topic modeling of k=4.

ES and 0.7% for ES-EN in terms of BLEU. The performance gain in terms of perplexity was 8% over the Moore-Lewis LM and 17% over the plain LM.

Further work includes the way to obtain the optimal k and the alternative way to set an appropriate (sentence-based) split of the corpus for given k and N (which was our assumption in Section 2.1). The latter may be related to the reason why the perplexity curve is quite complex with three local minima (at k=3, k=12, and k=19).

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of CNGL at Dublin City University.

References

[1] Moore, R.C., Lewis, W.: Intelligent selection of language model training data. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (2010) 220–224

[2] Blei, D., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR **3** (2003) 9931022

[3] Daumé III, H.: Frustratingly easy domain adaptation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (2007)

[4] Eidelman, V., Boyd-Graber, J., Resnik, P.: Topic models for dynamic translation model adaptation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (2012) 115–119

[5] Okita, T., Toral, A., van Genabith, J.: Topic modeling-based domain adaptation for system combination. In Proceedings of ML4HMT Workshop (collocated with COLING 2012) (2012)

[6] Du, L., Buntine, W., Johnson, M.: Topic segmentation with a structured topic model. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2013) 11

[7] Blei, D., Lafferty, J.: A correlated topic model of science. Anls of appl statistics **1** (2007) 1735

[8] Duh, K., Neubig, G., Sudoh, K., Tsukada, H.: Adaptation data selection using neural language models: Experiments in machine translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (2013) 678–683

[9] Stolcke, A.: SRILM – An extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing (2002) 901–904

[10] McCallum, A.K.: Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)

[11] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: 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 (2007) 177–180

[12] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method For Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02) (2002) 311–318

[13] Koehn, P.: Statistical significance tests for machine translation evaluation. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2004) (2004) 388–395