RESAMPLING AUXILIARY DATA FOR LANGUAGE MODEL ADAPTATION IN MACHINE TRANSLATION FOR SPEECH

Sameer Maskey, Abhinav Sethy

IBM T.J. Watson Research Center
New York, NY, 10598
{*smaskey, asethy*}*@us.ibm.com*

ABSTRACT

Performance of n-gram language models depends to a large extent on the amount of training text material available for building the models and the degree to which this text matches the domain of interest. The language modeling community is showing a growing interest in using large collections of auxiliary textual material to supplement sparse in-domain resources. One of the problems in using such auxiliary corpora is that they may differ significantly from the specific nature of the domain of interest. In this paper, we propose three different methods for adapting language models for a Speech to Speech (S2S) translation system when auxiliary corpora are of different genre and domain. The proposed methods are based on centroid similarity, n-gram ratios and resampled language models. We show how these methods can be used to select out of domain textual data such as newswire text to improve a S2S system. We were able to achieve an overall relative improvement of 3.8% in BLEU score over a baseline system that uses only in-domain conversational data.

Index Terms— Language Model Adaptation, Machine Translation, Domain Adaptation

1. INTRODUCTION

Most of the statistical machine translation systems [1], phrase based models [2] and syntax based system [3] require significant amount of data. We have seen in literature that usually Machine Translation (MT) engine's performance improves with the use of larger parallel corpus. But obtaining a large parallel corpus is difficult, costly and time consuming. On the other hand we have vast amount of non-parallel data on the web and from other sources such as LDC. One of the prevalent methods to use such non-parallel data in the translation system is to improve the language model (LM) component of MT using adaptation techniques.

In order to maximize the benefit from building language models from these generic corpora, we need to identify subsets of text relevant to the target application. In most cases we have a set of in-domain example sentences available to us which can be used in a semi-supervised fashion to identify the text relevant to the application of interest. The dominant theme in recent research literature for achieving this is the use of various rank-and-select schemes for identifying sentences from the large generic collection which match the in-domain data [4, 5]. An alternative to using the indomain data to identify the relevant text training material is to use the MT output of the test set to select data from the auxiliary out of domain sources. The selected data can then be used for adapting the indomain language model and the test set can be re-decoded with the adapted language model. The idea behind this two pass approach is to correct the implicit assumption that the test set is drawn from the same distribution as the training set [6]. We will show in our experiments that selecting relevant data with our proposed methods using indomain data gives significant improvements over the baseline system and that using the first pass MT output for the test set provides additional performance gains compared to using the indomain data.

We first present our data selection schemes based on text similarity, n-gram ratios and bagged estimates in Sections 3.1, 3.2 and 3.3 respectively. We then provide experimental results evaluating these three schemes in two different scenarios where we optimize LM by selecting data from auxiliary sources based on in-domain data vs. first pass MT output. We present our results in Section 4 and we conclude in Section 5. We next describe the related work on LM adaptation using auxiliary sources.

2. RELATED WORK

The central idea behind text data selection schemes for using auxiliary sources to build language models, has been to use a scoring function that measures the similarity of each observed sentence in the corpus to the domain of interest (in-domain) and assign an appropriate score. The subsequent step is to set a threshold in terms of this score or the number of top scoring sentences, usually done on a heldout data set, and use this threshold as a criterion in the data selection process. A dominant choice for a scoring function is in-domain model perplexity [4, 7] and variants involving comparison to a generic language model [8, 9]. A modified version of the BLEU metric which measures sentence similarity in machine translation has been proposed by Sarikaya [5] as a scoring function. Instead of explicit ranking and thresholding, it is also possible to design a classifier to Learn from Positive and Unlabeled examples (LPU) [10]. In LPU, a binary classifier is trained using a subset of the unlabeled set as the negative or noise set and the in-domain data as the positive set. The binary classifier is then used to relabel the sentences in the corpus. The classifier can then be iteratively refined by using a better and larger subset of the sentences labeled in each iteration. For text classification, SVM based classifiers are shown to give good classification performance with LPU [10].In [11][12] a relative entropy based subset selection scheme was proposed which tries to optimize the selection of the set as a whole in contrast to sentence selection by ranking.

In the broader context of statistical learning, the problem of selecting relevant data is akin to the classical problem of sample selection bias [6]. Resampling of training data for matching test and train distribution and correcting sample selection bias was used in [13] for better discriminative training of a maximum entropy classifier. In [14], resampling is used to select relevant auxiliary data for im-

proving a SVM classification model. Our work here does empirical evaluation of some of the techniques proposed in the related work and our new techniques which we will describe in the next section.

3. DATA SELECTION OF AUXILIARY DATA FOR LM ADAPTATION

In order to use auxiliary data to adapt LM we need to find sentences in auxiliary data that are similar to in-domain data or the first pass MT hypothesis. We present three data selection techniques below that scores each sentence in auxiliary data where we would like to show that scores can be used as rankings for the usability of sentence in improving LM. The first method is based on similarity of text to centroid vector of in-domain data. The second method is based on n-gram ratios and the third method uses resampled language models.

3.1. Method One (M1): Centroid Similarity

If you look at the sample of domain data one would notice that even though the sentences do not belong to a particular well defined topic as the ones defined in TDT4 they share some common topics that would occur in the conversation of doctors, interrogators, soldiers and inspectors. In order to take account of this domain we use a text similarity method to find sentences from auxiliary data that are not very far from the topics of the in domain data.

We first find a centroid vector of the in-domain data that would take account of all the topical terms of the in domain data. We compute centroid vector by computing TF.IDF(Term Frequency x Inverse Document Frequency)[15] and keeping only the terms that have TF.IDF scores higher than an empirically decided threshold. TF.IDF proposes that a term is significant if it rarely occurs in a corpus but occurs many times in a given document. In our experiments, TF and IDF is computed using the following equation.

$$TF(i,j) = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ the number of times the i_{th} term occurred in document j.

$$IDF(i) = log(\frac{N}{|d_j : t_i \in d_j|}) \tag{2}$$

where N is total number of documents in the corpus and the denominator is the total number of documents that contain the term i.

Now we compute TF*IDF weighted vectors for each sentence to compute its similarity with the centroid vector of the indomain data. Various text similarity metrics are available to identify similar sentences. We compute the similarity between sentences by computing cosine similarity. Cosine similarity between two documents or spans of text is defined as follows:

$$CentroidSimilarity(C,Y) = \frac{\vec{C} \cdot \vec{Y}}{||\vec{C}|| \cdot ||\vec{Y}||} \tag{3}$$

where $\vec{C}$ is the centroid vector for all of in-domain data and $\vec{Y}$ are word vectors for sentence Y, and $\vec{C} \cdot \vec{Y}$ is the dot product between them. Using the centroid similarity scores we rank all the sentences to obtain a first set of ranked set of sentences. The sentence with the highest centroid similarity score will be closest to the centroid of the in-domain data, i.e. it will have the most terms that describe the topics of the in-domain data.

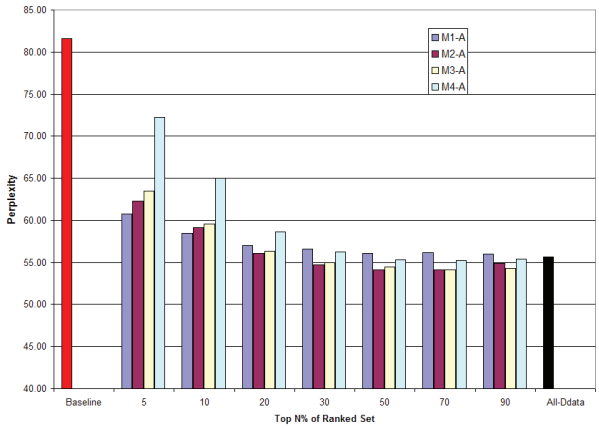

Fig. 1. Perplexity of LM Built with Selected Data at Various Thresholds

3.2. Method Two (M2): N-gram Ratio

Text similarity based methods for data selection identify sentences which are very similar and in many cases identical to the indomain set. However, we want the adaptation set to consist of sentences which provide additional n-gram coverage over the in-domain data. Based on our observations on the type of sentences that were being selected, we came up with a heuristic to ensure that sentences with previously unseen n-grams are selected while still maintaining an overall match with the indomain data. As a measure of match to in-domain data we want to select sentences which have a high probability with a lower order n-gram model (such as bigram or trigram model) built from in-domain data. At the same time, the selected sentences should introduce new higher order n-gram constructions which will get a low probability score from higher order indomain n-gram models. A weighted difference of the scores from the lower order and higher order n-gram model can thus be used as a measure to select sentences which not only match the in-domain data but also increase the n-gram coverage. In the experiments presented in this paper, we used a weighted difference of trigram language model and 4-gram scores $P_{3gr} - \lambda P_{4gr}$. We chose $\lambda = 0.1$ based on experiments on a smaller subset of data.

3.3. Method Three (M3): Ranking with Resampled language models

The motivation for the n-gram ratio (M2) method was to select sentences which boosted coverage while maintaining some degree of match to the in-domain set. An alternate viewpoint is to select data which when interpolated with the in-domain data helps lower the perplexity on some heldout set. Motivated by ensemble methods such as bagging and arcing[16], we resample the indomain set with a uniform distribution with replacement to create a bootstrap version of the indomain set T which has the same number of sentences as the in-domain set. Sentences that are excluded during resampling form the corresponding heldout set T_{-i}. We generate N such pairs of resampled indomain sets T_i and heldout sets T_{-i}. For each resampling, we build a language model $P_i(w)$ corresponding to T_i and a language model $P_{-i}(w)$ from T_{-i}. For the case of linear interpolation of models a good corpus for adaptation a would be such that $(\lambda)P_i(w) + (1-\lambda)P_a(w)$ matches the heldout model

$P_{-1}(w)$. We thus use the average difference between P_i and P_{-i} across various bagged estimates $i = 1..N$ to resample the data. The out of domain data is thus resampled or ranked using the weight $\sum_{i=1}^{N} P_{-i}(w) - \lambda P_i(w)$.

4. EXPERIMENTS AND RESULTS

We use a subset of the Malay to English parallel corpus provided by DARPA for the Transtac evaluation for our experiments. We refer to this corpus as the in-domain corpus C_d. The corpus consists of slightly above 100K parallel sentences for training with 732K words. We obtain a randomly sampled test and tuning set which have 1050 and 1340 sentences respectively. The sentences in the data is mostly conversational type such as "sure these are my keys", "my name is ahmad", "yes i have a receipt from the bank", and "thank you". The auxiliary corpus which is based on the text corpora available for HUB4 broadcast news evaluation consists of mostly newswire text and Broadcast conversations with sentences that are very different in domain and genre from the in-domain corpus. The LM training text in this auxiliary corpus consists of 335M words with approximately 16 million sentences from the following data sources: 1996 CSR Hub4 Language Model data, EARS BN03 closed captions, GALE Phase 2 Distillation GNG Evaluation Supplemental Multilingual data , Hub4 acoustic model training transcripts , TDT4 closed captions, TDT4 newswire, and GALE Broadcast Conversations and GALE Broadcast News. We collectively refer to our auxiliary corpus as C_a. Our goal is to select the data from C_a such that we can build a new LM for selected C_a and interpolate with current in domain model for a better performance, with the interpolation weight optimized on a heldout set.

In order to identify the best sentences from C_a, we first ranked all of the sentences in C_a using method M1, M2 and M3 which we have described in Section 3.1, 3.2, 3.3 respectively. After ranking the sentences in C_a with each method we selected top N% of sentences for N of 5, 10, 20, 30, 50, 70 and 90 and prepared individual sets for each value of N and the method. We obtained the sets M1.05,..., M1.90, M2.05,...,M2.90,M3.05...,M3.90. We wanted to compare the performance of our methods with a standard perplexity based method, which we call M4 for our experimental purposes. We ranked C_a using perplexity (M4) as well and again chose top N% of the sentences for values of N mentioned above to produce sets M4.05 to M4.90.

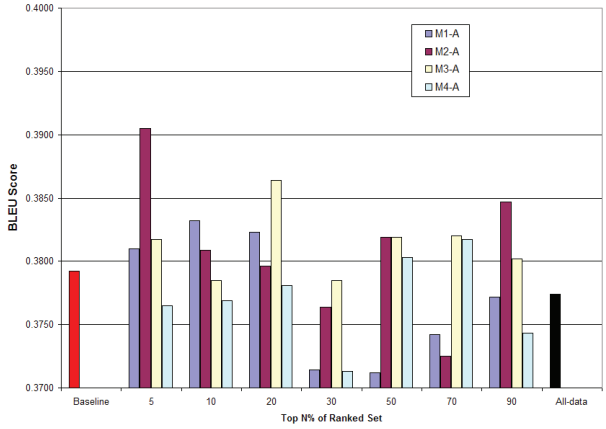

Fig. 2. Results with Selection based on In Domain Data

After we ranked the sentences in auxiliary corpus C_a and prepared these sets we wanted to find out which of the sets will produce the best LM $L_{M_i,N}$ which when interpolated with indomain LM (L_d) will result in a final LM that will improve the translation scores the most. One commonly used measure to test the performance of a language model is the language model perplexity on a heldout set.

In Figure 1 we observe a significant improvement on perplexity from the baseline model. Each of the adapted LMs corresponds to a bar in Figure 1. For example the language model built with 5% of data selected using method $M1$ ($LM_{M1,5}$) corresponds to the first bar after the baseline in the first cluster of bars in Figure 1. All these adapted LMs are built by interpolating the indomain LM with interpolation weights optimized on the heldout tuning set. We obtain 27.48 relative perplexity improvement with out best LM $L_{M2,70}$, while in general, perplexity improves with more auxiliary data for all of our methods. When we add all of the sentences in C_a we obtain perplexity of 55.62 which is 1.52 points lower than our best LM showing that adding all sentences may not be the best choice. We follow the convention used in Figure 1 to describe the other experimental results in this section.

Since our main motivation for this work was to improve MT by adapting LM, after testing LMs with perplexity we tested our LM by using them in MT experiments. Our MT engine is based on phrase translation model based closely to [2] with a stack decoder. We first built a baseline translation model using the parallel data of C_d with LM (LM_d) built from the target side (English) of the parallel corpus. For our experiments we replaced LM_d with the adapted LMs ($LM_{\{X=(M1,M2,M3,M4),Y=(5,10,20,30,50,70,90)\}}$). The results are shown in Figure 2.

We see that our best LM $LM_{M2,5}$ is higher than the baseline by 1.13 absolute BLEU points. Even though we had other LMs that use more data and had lower perplexity than $LM_{M2,5}$ we were able to get the most gain using only 5% of the data. On average over all the values of N our methods M1, M2 and M3 produce slightly better LMs than the LM based on sentences selected using perplexity. This shows that our data selection algorithm is better than the standard method of using perplexity for selecting data to adapt LMs. We also observe that our best adapted LM using 5% of auxiliary data is better than using LM that was trained with all of the auxiliary data by 1.31 BLEU points absolute. This further shows that our proposed method of data selection on auxiliary data to adapt LM is useful when the auxiliary corpus is of very different genre and domain. We observe in our results in Figure 2 and 1 that perplexity of LM may not correlate well with the MT performance so perplexity could be a poor choice for deciding which LM to use for MT.

In place of the in-domain language model, selecting data according to the test set we are translating should probably give more gains. But since we cannot assume to have the translation of the test set we can at best do an adaptation based on the "potential translation". In order to adapt the LM more to the test set we first get the first pass decode of the translation, i.e. given a Malay test set S_t we get it's first pass English decode T_t. Given the first pass decode we use methods to rank all the sentences of the C_a corpus. Finding sentences that has similar n-grams to the current first pass decode could produce better ranking of the sentences because instead of optimizing the LM to the training data, we are optimizing LM to better estimate the n-gram counts of the pseudo test set. We see in Figure 3 that the model $L_{M2,10}$ does better than the baseline by 1.46 bleu points when it is tuned on the first pass decode, which is even higher than the best BLEU scores we obtained in Figure 2 by 0.33 BLEU points. In general we see that for most if the adapted LMs

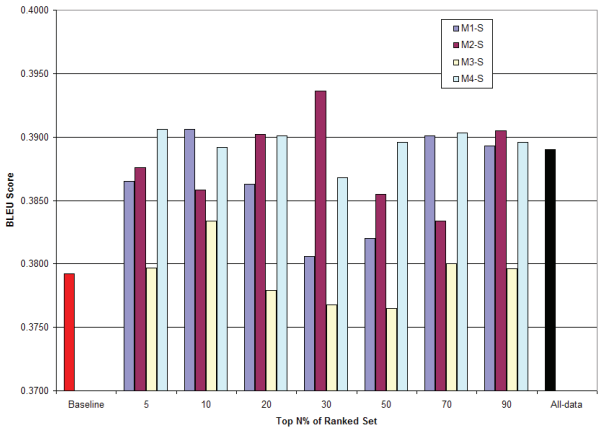

Fig. 3. Results with Selection based on First Pass Decode

in Figure 3 we see gain over the LMs in Figure 3. This shows that in general adapting LM based on the first pass decode is better than adapting LM to in-domain data when we are using a large amount of auxiliary data source of different genre and domain. It should be noted that the baseline corresponding to using all of the data is different from Figure 3 since the interpolation weights between the indomain and out-of-domain LMs was reoptimized on the first pass hypothesis.

In our experiments above where we tested the three methods we proposed (M1, M2 and M3) for ranking auxiliary sentences and used the ranking information to adapt the LM for MT we saw that our method based on n-gram ratios performed better than a commonly used method based on perplexity only. We also saw that we can improve our LM further if we adapt it based on the first pass decoder output from MT engine.

5. CONCLUSION

In this paper we presented an empirical evaluation of three methods we propose for adapting language models for Speech to Speech translation system by selecting relevant sentences from auxiliary out of domain data. We report results in a semi supervised setting where a small indomain set was used to seed the selection and a two pass scenario where the first pass decode of the MT system is used in place of the in-domain data. In both the semi supervised and the two pass experiments we observed significant improvements in BLEU score. The method that performed best in our experiments with in-domain data is based on a weighted difference of sentence likelihood between lower order and higher order n-gram models. The second best performing method uses resampling to generate multiple in-domain and heldout sets. In our experiments where the first pass output was used for data selection, all method performed very closely although the n-gram ratio method still gave the best BLEU score.

We plan to extend the n-gram ratio method to do subset selection instead of rank based selection as described in [11, 12]. This can be achieved by including the ngram ratio as an additional objective function to minimize jointly with relative entropy. Another direction of work under investigation is resampling the parallel corpus based on the first pass decode and to subsequently use the resampled data for building interpolated translation tables.

6. ACKNOWLEDGMENT

This work is in part supported by the US DARPA under the TransTac (Spoken Language Communication and Translation System for Tactical Use) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

7. REFERENCES

[1] Peter E Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, pp. 263–311, 1993.

[2] Franz Josef Och and Daniel Marcu, "Statistical phrase-based translation," 2003, pp. 127–133.

[3] David Chiang, "A hierarchical phrase-based model for statistical machine translation," in *In ACL*, 2005, pp. 263–270.

[4] Tim Ng, Mari Ostendorf, Mei-Yuh Hwang, Manhung Siu, Ivan Bulyko, and Xin Lei, "Web-data augmented language model for Mandarin speech recognition," in *Proceedings of ICASSP*, 2005.

[5] Ruhi Sarikaya, Agustin Gravano, and Yuqing Gao, "Rapid language model development using external resources for new spoken dialog domains," in *Proceedings of ICASSP*, 2005.

[6] James Heckman, "Sample selection bias as a specification error," *Econometrica*, 1979.

[7] Teruhisa Misu and Tatsuya Kawahara, "A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts," in *Proceedings of ICSLP*, 2006.

[8] Abhinav Sethy, Panayiotis Georgiou, and Shrikanth Narayanan, "Building topic specific language models from web-data using competitive models," in *Proceedings of Eurospeech*, 2005.

[9] Karl Weilhammer, Matthew N Stuttlem, and Steve Young, "Bootstrapping language models for dialogue systems," in *Proceedings of ICSLP*, 2006.

[10] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip Yu, "Building text classifiers using positive and unlabeled examples," in *Proceedings of ICDM*, 2003.

[11] Abhinav Sethy, Panayiotis G. Georgiou, and Shrikanth Narayanan, "Text data acquisition for domain-specific language models," in *Proceedings of EMNLP*, 2006.

[12] Abhinav Sethy, Panayiotis G. Georgiou, Bhuvana Ramabhadran, and Shrikanth Narayanan, "An iterative relative entropy minimization based data selection approach for n-gram model adaptation," *IEEE Transactions on Audio, Speech and Language Processing*, To be published.

[13] Steffen Bickel, Michael Brckner, , and Tobias Scheffer, "Discriminative learning for differing training and test distributions," in *Proceedings of ICML*, 2007.

[14] P Wu and T G Dietterich, "Improving svm accuracy by training on auxiliary data sources," in *Proceedings of ICML*, 2004.

[15] Gerard Salton and Chris Buckley, "Term weighting approaches in automatic text retrieval," Tech. Rep., Ithaca, NY, USA, 1987.

[16] L. Breiman, "Bias, variance, and arcing classifiers," Tech. Rep., 1996.