ADVANCES IN SYNTAX-BASED MALAY-ENGLISH SPEECH TRANSLATION

Bing Xiang, Bowen Zhou, Martin Čmejrek

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{bxiang, zhou, martin.cmejrek}@us.ibm.com

ABSTRACT

In this paper, we present advanced techniques that improved the performance of IBM Malay-English speech translation system significantly. During this work, we generated linguistics-driven hierarchical rules to enhance the formal syntax-based translation model; designed an active learning approach with bi-directional translations that outperformed unsupervised training; utilized translation direction information in parallel training corpus to build direction-specific interpolated language models for machine translation. There is 20% relative improvement achieved in the translation performance through all these techniques. A state-of-the-art Malay speech recognition system was also established as one of the crucial modules in the rapidly developed Malay-English speech translation.

Index Terms— Machine Translation, Speech Recognition, Active Learning

1. INTRODUCTION

Speech translation has become an active area in recent years. It covers both automatic speech recognition (ASR) and machine translation (MT) and calls for innovative ideas from both fields. In this work, we report significant progress we have achieved during the rapid development of speech translation between a low-resource language pairs, Malay and English. We will cover the following topics: linguistics-driven hierarchical rules to enhance formal syntax-based translation model; active learning with bi-directional translations; direction specific interpolated language models.

Syntax-based translation model has drawn much attention in machine translation community recently. It incorporates both phrase-based translation structure and synchronous context-free grammar (SCFG), and has shown promising progress in various translation tasks. Syntax-based translation models can be further categorized into two classes: formal syntax-based and linguistical syntax-based. The former automatically extracts synchronous grammar from parallel corpus without explicit usage of linguistic knowledge (e.g. [1]), while the latter relies on the syntactic parsing information on at least one side of the parallel corpus (e.g. [2][3]). Some approaches combining both classes have been proposed recently. In [4], a prior derivation model is incorporated using linguistically syntactic parsing to improve the performance of formal syntax-based translation model. In this work, with a similar motivation as in [4], we enlarge the data-driven hierarchical rule set with a set of linguistics-driven rules to achieve better coverage of rules and also more accurate long-distance reordering. Those additional rules are automatically created based on part-of-speech (POS) tagging on one side of the parallel corpus and catch some of the canonicalized reordering phenomena in case they are missed by the blindly data-driven hierarchical rules, especially when the training data is rather limited.

In this work, we also propose a new active learning approach for statistical machine translation. Various unsupervised and semi-supervised training techniques have been proposed for speech recognition and machine translation in recent years. For example, unsupervised training is shown as an effective approach in [5][6] for speech recognition, and in [7] for machine translation. An approach that combines active and semi-supervised learning was also proposed in [8] for spoken language understanding. As a continuation of the work in [7], we extend the approach to active learning here. We first translate monolingual data with the baseline system, then select sentences that are the most difficult to translate by current system so that human can make corrections. Such approach can be efficient and effective for rapid development of translation system on new or low-resource languages.

It is known that language model (LM) is an important component in both speech recognition and machine translation. The approach of interpolating multiple language models built from different domain data has been widely applied. In this work, we point out that the translation direction information in parallel corpus is crucial and can be utilized to improve the language model.

The rest of the paper is organized as follows: Section 2 presents formal syntax-based baseline model and linguistics-driven hierarchical rules. Section 3 describes the active learning approach. Section 4 briefly mentions the direction-specific interpolated language models. Section 5 introduces the development of a state-of-the-art Malay speech recognition system. Section 6 reports extensive experimental results obtained in Malay-English translation. The paper ends with some conclusions and future work discussion in Section 7.

2. COMBINING LINGUISITIC KNOWLEDGE WITH FORMAL SYNTAX

In this section we first describe our baseline, a formal syntax-based MT system, then present the approach of combining formal syntax with linguistic knowledge.

2.1. Formal Syntax-based Translation

Our baseline contains a formal syntax-based translation model [4]. It utilizes SCFG, with each synchronous production, i.e. rule, rewriting a non-terminal into a pair of strings, source string S and target string T, in both languages. Each string can contain both terminals and non-terminals, under the constraint that there is one-to-one correspondence between non-terminals at the source and target side. A unified symbol X is used for all non-terminals in the rule set.

$$X \rightarrow < S, T, \sim >, \tag{1}$$

where $\sim$ is the link between non-terminals in S and T. A glue rule is embedded with decoder to allow sequential concatenation of sub-

translations.

$$X \rightarrow < X_1X_2, X_1X_2 > . \quad (2)$$

During the system building, we start from a sentence-aligned parallel training corpus and generate word alignments with GIZA++ [9] based on IBM Model 1-4 and hidden Markov model. Then we extract phrase pairs based on the word alignments and some symmetrization heuristics [9]. A phrase table is built upon them with the probabilities estimated based on relative frequency. Abstract rules are extracted based on generalization of phrase pairs, similar to [1]. Each abstract rule has one or two non-terminals. A set of pruning techniques are applied to control the size of rule set. The features used in the decoder include phrase translation probabilities and lexical probabilities in both directions, language model, word counts, rule counts, glue rule penalty and abstraction penalty[4]. The decoding weights are optimized to maximize BLEU scores [10]. A 4-gram language model is trained with Kneser-Ney smoothing [11] and used in the decoder.

2.2. Linguistics-Driven Rules

We tried to improve our baseline translation system by adjusting it to some general systemic differences between English and Malay to compensate for insufficient observations, such as word order of noun phrases, and forming questions.

As in English, the basic word order in Malay is Subject Verb Object. However, there are several differences in constituting complex noun phrases. Adjective pre-modifications in English are usually ordered from the most specific to the most general, while in Malay they appear as post-modifications in reverted word order, e.g. *dark blue color* $\leftrightarrow$ *warna(color) biru(blue) gelap(dark)*. Demonstrative and possesive pronouns follow the noun, e.g. *this car* $\leftrightarrow$ *kereta(car) ini(this)* or *your car* $\leftrightarrow$ *kereta(car) awak(your)*.

While analyzing the English to Malay translation output, we observed that noun phrases that had to be covered by combinations of short 1-1 phrase pairs often use the English word order. In order to encourage these word reordering and also to increase the word coverage, we decided to add high probability rules such as

$$X \rightarrow < X_1N_e, N_mX_1 > \quad (3)$$

for all English-Malay noun translation pairs $N_e \leftrightarrow N_m$, and rules such as

$$X \rightarrow < J_eX_1, X_1J_m > \quad (4)$$

for all adjective translation pairs $J_e \leftrightarrow J_m$.

We also tried to improve the performance on general questions, such as *Do you like apples...* by inserting rules of form

$$X \rightarrow < do/does X_1V_eX_2, X_1V_mX_2lah > \quad (5)$$

for all verb translation pairs $V_e \leftrightarrow V_m$.

Since we had no POS annotated data for Malay available, in order to obtain the necessary POS specific translation pairs, we processed the English part of the parallel corpus by a syntactic parser [12], and extracted lists of English nouns, adjectives, and verbs. Then we found their translations in publicly available online dictionaries, using all one-word translation alternatives.

3. ACTIVE LEARNING

In this work, we also propose a combination of active learning with unsupervised training to alleviate the low resource problem of sentence-aligned parallel training data. Parallel corpus is an essential resource for developing machine translation systems. However, it is typically expensive and time-consuming to collect such parallel corpus when we need to develop a system for new language pairs.

Similar with our previous work in [7], we can take advantage of available monolingual data in one of the languages. First we translate those monolingual data using the baseline system in both directions. The translation hypotheses are selected from re-ranked N-best list based on confidence scores as in [7]. Then we select the sentences that current system has the biggest problem with for human to correct. For those sentences with confidence scores higher than certain threshold, we add them directly into the original parallel corpus. While for those with lowest confidence scores below a certain threshold, we present them automatically to human translators to check the quality and make corrections before adding them to the parallel corpus and retraining the system. The procedure can be iterative. In this way, we can quickly establish an MT system with good performance on the new language pairs by always actively expanding data that is mostly needed by the translation models.

Assume we need to build an MT system between language E and F, and we have a parallel corpus C_1 to start with. In the meantime, we have a large corpus E_1 in language E that is relevant to our target task.

The algorithm is as following: Start with iteration $i = 1$,

1. Build a two-direction translation system S_i using parallel corpus C_i for $E \rightarrow F$ and $F \rightarrow E$.
2. Use $E \rightarrow F$ model of system S_i to translate sentences in E_i to hypotheses F_i.
3. Use $F \rightarrow E$ model of system S_i to translate F_i to hypotheses E_i'.
4. Measure the similarity or distance between sentences in E_i and E_i' using some scoring metrics. In this work we use BLEU.
5. Rank the similarity scores of all sentence pairs in E_i and F_i.
6. Pick sentence pairs $E_l \leftrightarrow F_l$ which have lowest similarity scores, and present them to human translators to correct the translations and add to parallel corpus.
7. Add also sentence pairs $E_h \leftrightarrow F_h$ with highest confidence scores into the parallel corpus for next iteration, so $C_{i+1} = C_i + E_l + F_l + E_h + F_h$.
8. Make $E_{i+1} = E_i - E_l - E_h$ and $i = i + 1$, go back to step 1, until certain stop criterion is met.

The advantage of the active learning approach is that at each iteration, we pick a subset of the corpus which has lowest scores, when E translates into F and then translates back to E, for human checking and correction. This approach tries to maximize the human correction effectiveness. In the meantime, in each iteration we also add sentence pairs with high confidence scores to current corpus. In this way, the phrase coverage can be quickly improved.

4. INTERPOLATED LM

Interpolated language models have been widely used as a way to adapt the model towards different domain. Here we propose an approach to take into consideration of the direction information in the parallel training data for speech translation. The training data for target languages is partitioned into two subsets based on the translation direction ($E \rightarrow F$ or $F \rightarrow E$). For tasks like those in the DARPA TRANSTAC program, there are mainly two types of speaker roles:

interviewer who asks questions, and respondent who answers questions. It is understandable that the speech content from each direction varies significantly. This information can be utilized to build a better language model for specific test scenarios, e.g. translation of respondent speech. There is also a way to bias the translation model instead of language model, by simply duplicating multiple times the data from certain subset in the parallel corpus. We will compare these two approaches in Section 6 later.

5. MALAY SPEECH RECOGNITION

In this section we briefly describe the development of a state-of-art Malay speech recognition system.

Our Malay acoustic training data consists of around 90 hours of speech collected under the DARPA TRANSTAC program. Every 10 ms a 24-dimensional MFCC feature vector is computed and then mean normalized. Sequences of 9 vectors are then stacked together leading to a 216-dimensional new feature vector. This new feature space is finally reduced to 40 dimensions with a combination of linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT).

We built context-dependent quinphone models with 90K Gaussians and 35 phonemes. The models are trained under either maximum likelihood (ML) or minimum phone error (MPE) [13] criterion, with an fMPE [14] feature transform applied to the latter. Online speaker adaptation with vocal tract length normalization (VTLN) and feature space maximum likelihood regression (fMLLR) are utilized to further improve the performance. A statistical 3-gram language model with Kneser-Ney smoothing [11] is trained with around 100K Malay sentences.

Table 1 shows the roadmap of the development of our Malay ASR system. The word error rate (WER) is measured on a 3-hour Malay test set randomly selected from the acoustic data and excluded from the training. Three types of data are included in the test set, from 1.5-way (t_1), 2-way monolingual (t_2) and 2-way bilingual (t_3) corpora respectively, as shown in Table 1.

Train	Iter	t1	t2	t3	All
ML	1	11.98	33.33	18.36	21.72
ML	2	11.32	31.09	18.25	20.74
+VTLN	1	11.03	28.64	15.93	19.00
+fMLLR	1	10.58	25.84	14.92	17.52
+MPE+fMPE	4	8.92	22.19	11.95	14.71

Table 1. ASR results on the Malay test set

In the second iteration of ML training, we used the previous iteration model to align the audio data against the transcripts to generate better alignment, which helped by 1% in WER. VTLN and fMLLR together reduced the WER by another 3% absolute. Finally discriminative training in feature and model space, i.e. fMPE and MPE, reduced the WER down to 14.71%. The improvement is consistent across three different types of data in the test set.

6. EXPERIMENTAL RESULTS

In this section, we report the results we obtained on a set of experiments conducted in translations between English and Malay. The translation model in the baseline system is trained with 100K parallel sentences provided under the TRANSTAC program. 4-gram language models are trained with the target side data from the parallel corpus. The system weights are tuned on a development set with around 1K sentences. There are 1340 sentences in Malay-to-English test set and 1050 sentences in English-to-Malay test set. The experiments in the first three subsections below are conducted on English-to-Malay translation, with speech reference as input. The last subsection reports speech-to-text (S2T) translation results in both directions. All test sets have one set of human-annotated MT references.

6.1. Interpolated LM

As mentioned earlier, there are two types of directional data in the parallel training corpus. The distribution of the translation directions are shown in Table 2, where E stands for English and M for Malay. Separate Malay language models are trained with each subset of data. Then interpolation weights are tuned on the development set to minimize the perplexity.

Direction	Sentences	Weights	Perplexity
$E \rightarrow M$	6K	0.64	310.22
$M \rightarrow E$	94K	0.36	493.05
Interpolated	100K		212.43

Table 2. Perplexities from different subsets

Since in our English-to-Malay translation task, the English input is dominated by interviewer speech, the component trained with E-to-M data was assigned the much larger weight than that on the other direction, even though the data is little among the whole training corpus.

System	Interp LM	Dup Train	Sentences	BLEU
Baseline	No	No	100K	19.23
System I	No	Yes	130K	20.19
System II	Yes	No	100K	20.80
System III	Yes	Yes	130K	21.02

Table 3. Interpolated LM vs. duplication of parallel data

The translation results are shown in Table 3. We compare the effect of interpolated LM with the duplication of E-to-M data by 6 times. We can see that compared to the baseline, duplication of parallel data achieved 1 point of gain in BLEU. While using interpolated LM only, there is a larger gain obtained with 1.6% absolute. When using both approaches, another 0.2% gain is achieved. These results show that both techniques are helpful and interpolated LM provides relatively larger benefit.

6.2. Active Learning

In Table 4, we compare the results of active learning and unsupervised training. We translate 90K English sentences with system III in Table 2 first using unsupervised technique back and forth in both directions as described in Section 3. Then we sort the sentence pairs according to BLEU-based similarity scores. We added different amount of sentences, starting from those with highest scores, to the original corpus, as in row 2 to 4 in Table 4. There is little change when adding the top 20K sentences. With 70K sentence pairs, we have with more phrase pairs and abstract rules in the model and obtained 0.4% gain. Adding all 90K sentences, there is no more improvement. The bad translation from the bottom sentences hurts the performance instead.

Orig	Unsupervised	Active	Phrases	Rules	BLEU
100K	0	0	1.8M	3.1M	21.02
100K	20K	0	2.1M	4.0M	21.12
100K	70K	0	2.9M	6.8M	21.43
100K	90K	0	3.7M	8.8M	21.21
100K	70K	20K	3.8M	8.9M	22.17
100K	70K	20K	3.8M	4.0M	22.43

Table 4. Active learning vs. unsupervised training

We picked the bottom 20K sentence pairs that have the lowest scores for human to correct, then added the correct translation into the original corpus along with the 70K sentences selected with unsupervised training. There is 0.7% absolute gain from adding the active learning data. Since when generalizing the phrase pairs obtained from unsupervised data, there could be more noise added into the abstract rule set, we try to exclude those abstract rules obtained from unsupervised data, as in the last row in Table 4, another 0.3% gain was obtained in BLEU.

6.3. Linguistics-Driven Rules

In Table 5, we show the effect of adding 33K linguistics-driven hierarchical rules to the original data-driven abstract rule set. There is 0.4% gain obtained purely from these additional rules, even though the number of extra rules is less than 1% of the original rule set. This shows that linguistic rules have much higher quality and more focused than the data-driven rules. They also provided extra rules that the original rule set couldn't cover due to limited amount of training data.

System	Phrases	Rules	BLEU
Formal Syntax	3.8M	4.0M	22.43
+ Linguistic rules	3.8M	4.1M	22.84

Table 5. Linguistics-driven rules

With the three techniques mentioned above, we eventually increased the BLEU score from 19.23% to 22.84%, around 20% relative improvement, which is significant.

6.4. S2T

We also show the improvement in S2T in both translation directions in Table 6, where the Malay ASR WER is around 15% and English around 11%. The final model utilized all techniques described above. Significant improvement has been achieved in both directions for S2T task, similar with the results obtained on the translation of speech references.

Direction	System	S2T
$M \rightarrow E$	Baseline	30.44
$M \rightarrow E$	Final	33.65
$E \rightarrow M$	Baseline	16.69
$E \rightarrow M$	Final	20.91

Table 6. Speech translation results

7. CONCLUSIONS AND FUTURE WORK

We have described severval advanced and effective approaches that improved the performance of our Malay-English speech translation system significantly. With the help of linguistics-driven rules, active learning, direction specific interpolated language models, the BLEU score was increased by around 20% relative. We also established a state-of-the-art Malay speech recognition system for Malay-English speech translation. Our next work will focus on more intelligent rule extraction, selection and filtering techniques. A combination of linguistic knowledge and formal syntax will continue to be an interesting topic for machine translation.

8. ACKNOWLEDGEMENT

This work was supported by the DARPA TRANSTAC program.

9. REFERENCES

[1] D. Chiang, "Hierarchical phrase-based translation," in *Computational Linguistics*, 2007, pp. 201–228.

[2] K. Yamada and K. Knight, "A syntax-based statistical translation model," in *Proc. ACL*, 2001, pp. 523–530.

[3] M. Galley, M. Hopkins, K. Knight, and D. Marcu, "What's in a translation rule?" in *Proc. HLT/NAACL*, May 2004.

[4] B. Zhou, B. Xiang, X. Zhu, and Y. Gao, "Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels," in *Proc. Second ACL Workshop on Syntax and Structure in Statistical Translation*, 2008, pp. 19–27.

[5] L. Lamel, J. L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Proc. ICASSP*, 2002, pp. 877–880.

[6] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *Proc. ICASSP*, 2006, pp. 1056–1059.

[7] B. Xiang, Y. Deng, and Y. Gao, "Unsupervised training for Farsi-English speech-to-speech translation," in *Proc. ICASSP*, 2008, pp. 4977–4980.

[8] T. Gokhan, D. Hakkani-Tur, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," in *Speech Communication*, 2005, pp. 171–186.

[9] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," in *Computational Linguistics*, vol. 29, 2003, pp. 9–51.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, July 2002, pp. 311–318.

[11] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. ICASSP*, 1995, pp. 181–184.

[12] E. Charniak, "A maximum-entropy-inspired parser," in *Tech. Report CS-99-12, Brown University*, 1999.

[13] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. 105–108.

[14] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: discriminatively trained features for speech recognition," in *Proc. ICASSP*, 2005, pp. 961–964.