POSTECH Machine Translation System for IWSLT 2008 Evaluation Campaign

Jonghoon Lee and Gary Geunbae Lee

Department of Computer Science and Engineering
Pohang University of Science and Technology
{jh21983, gblee}@postech.ac.kr

Abstract

In this paper, we describe POSTECH system for IWSLT 2008 evaluation campaign. The system is based on phrase based statistical machine translation. We set up a baseline system using well known freely available software. A preprocessing method and a language modeling method have been applied to the baseline system in order to improve machine translation quality. The preprocessing method is to identify and remove useless tokens in source texts. And the language modeling method models phrase level n-gram. We have participated in the BTEC tasks to see the effects of our methods.

1. Introduction

In this paper, we describe our MT system for IWSLT 2008 evaluation campaign. We have been developing a statistical machine translation system based on Moses system [1] which is an open source phrase based machine translation system.

Our ongoing research topics are preprocessing based on morphological information and advanced language modeling to model longer history effectively. We have applied our findings from experiences of Korean-English translation into translating some other language pairs.

We have participated in the three BTEC tasks: Arabic to English, Chinese to English, and Chinese to Spanish. Although we have almost no knowledge and experiences in Arabic, Chinese, and Spanish, a language independent characteristic of SMT techniques made the participation possible.

The following section describes our baseline system and statistics of supplied data. And section 3 describes two methods applied to improve the baseline system. Section 4 contains evaluation results and some discussions. Section 5 concludes this paper.

2. Baseline system

We have used Moses system in order to build the phrase-based SMT systems for IWSLT 2008 evaluation campaign. Phrase-based approaches to SMT usually use a number of feature functions those are combined in a log-linear model. We used the following features those are presented by the default setting of Moses system.

- Source to target and target to source phrase translation probabilities
- Source to target and target to source word translation probabilities (lexical weightings)
- Phrase penalty (a constant by default)
- Word penalty
- Distance based distortion model

A target language model was used in addition to the features. We have used the SRILM toolkit [2] in order to build the target language model. The weights for the features are optimized by minimum error rate training [3] which maximizes BLEU score. We have used only IWSLT 2008 train and development data for training translation and language model. The corpus statistics are summarized in table 1.

Table 1. Corpus statistics of supplied data for Arabic-English, Chinese-English, and Chinese-Spanish tasks: Word counts and vocabulary sizes are measured after preprocessing steps

		Arabic	Chinese	English	Spanish
Train	Sent.		19972		
	Word	150303	171591	189558	185527
	Vcb.	14854	8428	8170	10995
Dev1	Sent.	506	506	506*16	
	Word	2865	3354	61176	
	Vcb.	1102	880	983	
Dev2	Sent.	500	500	500*16	
	Word	3040	3449	61615	
	Vcb.	1180	920	979	
Dev3	Sent.	506	506	506*16	506*16
	Word	2918	3767	62690	60501
	Vcb.	1174	931	997	1151
Dev4	Sent.	489	489	489*7	
	Word	4825	5715	46042	
	Vcb.	1473	1143	1157	
Dev5	Sent.	500	500	500*7	
	Word	5341	6066	51874	
	Vcb.	1797	1339	1354	
Dev6	Sent.	489	489	489*6	
	Word	2757	3169	22366	
	Vcb.	1119	881	924	
Test	Sent.	507	507		
	Word	2955	2808		
	Vcb.	1139	885		

3. Our methods to improve

3.1. Deleting useless tokens

Each language has its unique word formation strategy and morphological structure. In machine translation, some morphological phenomena observed in a source language could not be found in a target language and vice versa. The difference between source and target language could make some *useless tokens* in statistical machine translation. We define the term *useless token* as follows:

- In parallel texts, if a token does not have

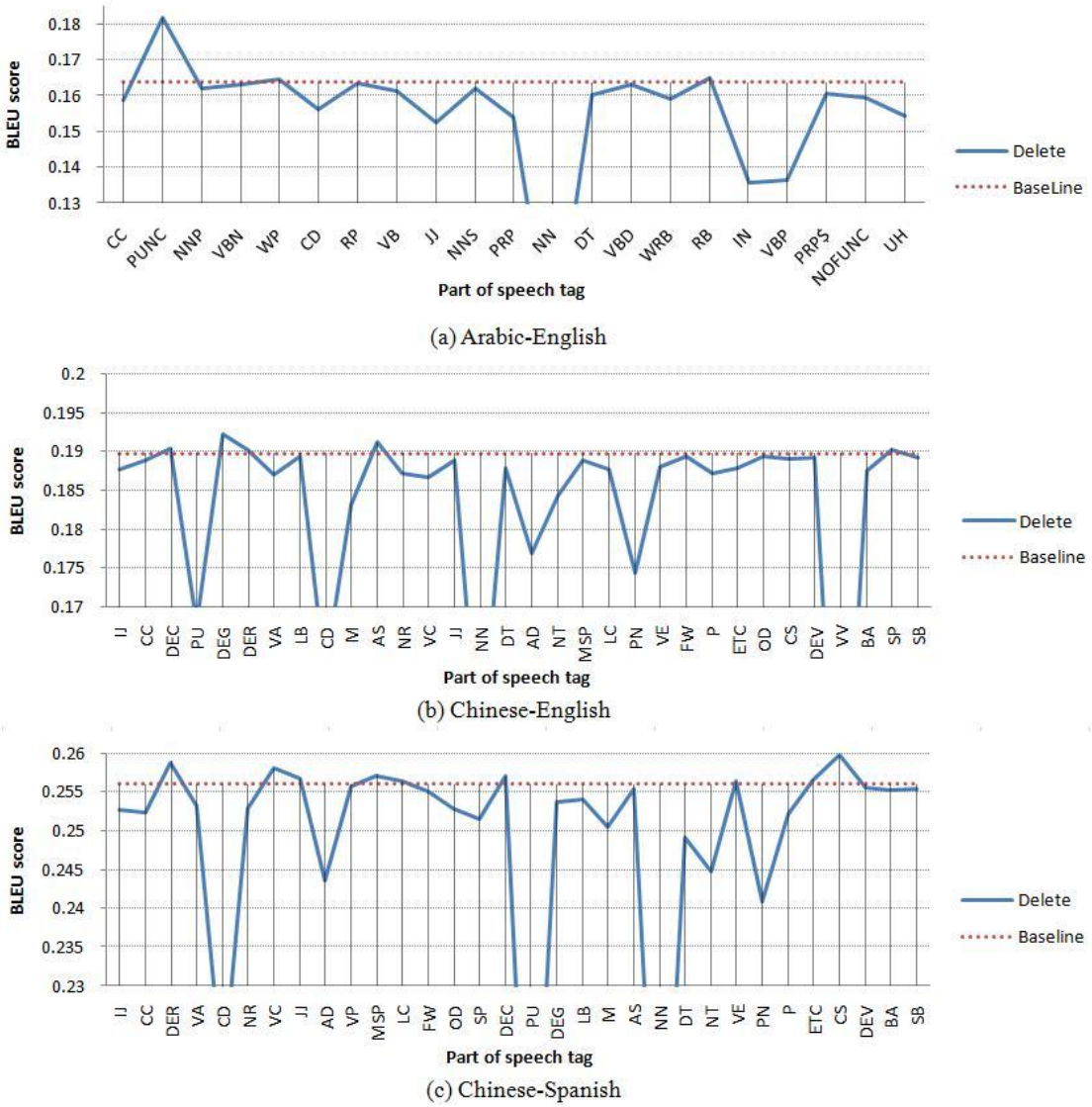

Figure 1. Deletion test results

corresponding tokens of same meaning or function in the opposite side text, the token is *useless*.

The useless words should be aligned with NULL position because they have no proper words to be matched with, by the definition. However, we observed that the useless words are usually aligned incorrectly in other experiments when we use GIZA++ [4] to get the alignment. These erroneous alignments should be refined or removed in order to improve machine translation quality. Our approach to the problem is to delete the useless words before word alignment stage to prevent the incorrect word alignment caused by the useless words.

In order to precisely identify the useless words, careful comparison between source and target languages based on linguistic insight is necessary. However, the comparison is not available because the authors do not have any knowledge in the source languages of BTEC tasks: Chinese and Arabic. As an alternative, a series of deletion tests have been performed to identify the useless tokens.

A deletion test is a very simple empirical method. For each candidate, the deletion test is done by training and testing a SMT system after deleting the candidate. We decide that the candidate is useless if the deletion test improves machine translation quality in terms of BLEU score [5]. This decision may not always agree with the definition of useless tokens. However, the performance improvement is a strong evidence of useless tokens. Assuming that the useless words are distributed in several parts of speech (POS), we performed the deletion test for each POS tag because performing the test for all vocabulary is too time consuming.

Figure 1 shows the results of the deletion tests for the three BTEC language pairs. The deletion tests have been carried out by using all the development corpora as a development corpus, i.e., dev1 to dev6 are merged for Arabic-English and Chinese-Spanish pairs. Arabic texts have been tokenized and labeled with POS tags using Arabic SVM Tools [6]. Chinese POS tagging has been done by Stanford parser [7] on the given tokenizing.

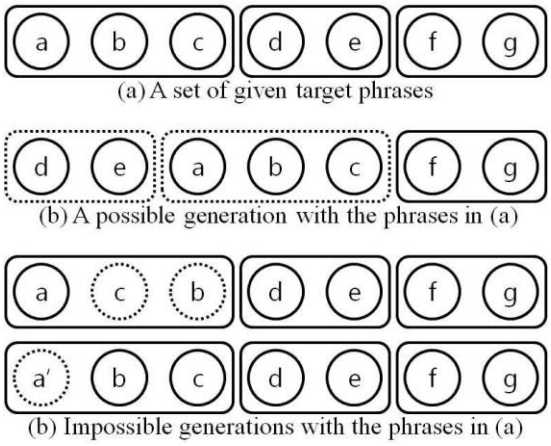

Figure 2. Limitations on generating target sentences in phrase based framework.

We classified the POS tags which result in a point higher than the baseline into useless; translation quality is improved by *deleting* those tokens. Roughly speaking, the BLEU scores in figure 1 represent degree of uselessness.

In Arabic-English task, only one POS tag, 'PUNC' has been classified into useless POS tag, i.e., the tag mainly consists of useless words. In fact, it means that no Arabic POS are useless because the PUNC tag marks for punctuations. However, we expect that more useless POS tags can be found if we perform the test with more refined tag set and tokenization. Arabic may contain functional morphemes which are not observed in English; Arabic is a morphologically rich language.

In Chinese-English task, we have found five useless POS: DEG, DEC, DER, AS, and SP. However, the changes of BLEU score are too small to classify the POS into useless, except DEG. The DEG is a tag for genitive and associative markers. Although English has a genitive marker *'s*, it is not frequently observed in English texts of BTEC corpus. Therefore, to classify DEG tag into useless in Chinese-English translation is reasonable; it satisfies the definition of the useless token.

In Chinese-Spanish task, more useless tags are observed than Chinese-English task. The useless tags for Chinese-Spanish translation are ER, VC, JJ, MSP, LC, DEC, VE, ETC, and CS. However, most of them have too weak empirical evidence to classify them into useless. DER and CS show the biggest improvement. Unfortunately, however, we cannot confirm whether the two tags are really useless, due to the absence of linguistic knowledge in Spanish.

3.2. Phrase level Language Model

Moving from word-based to phrase-based machine translation [8], [9] significantly improved translation quality by capturing local reordering within aligned phrase pairs. In this framework, generating target sentences is not done at a single-word level. It never occurs to change some words or their ordering in a given phrase, as described in figure 2. We would call reordering and selecting at the single word level 'an inner-phrase decision' and doing so at phrase level 'an inter-phrase decision.'

In word based systems, selecting and reordering target words for fluency were originally language model's role. During the decoding process of phrase based SMT systems, however, the inner-phrase decision is not controlled by the word based language model. Actually, two important roles of the language model in Moses decoder are future cost scoring and inter-phrase reordering. Therefore, each phrase pair can be treated as an atomic unit for language models as well as translation models. We have been developing a language modeling method that models target language at phrase level for phrase based machine translation systems in order to strengthen inter-phrase decisions during the decoding process.

Building phrase based language model can be decomposed into two sub-problems: identifying phrase level vocabulary and building a language model within the vocabulary. We have noticed that the translations of the phrase-based machine translation systems are generated by combining phrase pairs pre-defined in a translation model, i.e., phrase table. Target phrases in the phrase table are enough to cover all possible decoder output. By using the target phrases in the phrase table, the first problem has been solved.

Our approach to the second problem is to use traditional back-off n-gram modeling methods. Modeling phrase n-gram dependency is a conceptually same method as modeling word n-gram dependency. However, extracting phrase n-gram counts is slightly different from extracting word n-gram counts. A sentence has a unique tokenization at word level; each word is a fixed unit that does not overlap with other words. On the other hands, tokenizing a sentence at phrase level generates a lot of candidates; each word can be contained in more than one phrases. We define the count of phrase n-gram for a sentence as maximum count that can be observed in a candidate tokenization of the sentence. We get the counts for all possible phrase n-gram sequences; the sentence can contribute more than one count for lots of phrase n-grams. Phrase based n-gram model is built by SRILM toolkit from the n-gram count.

The phrase based n-gram definitely suffers from relatively severe data sparseness because the phrase level data are sparser than the word level data. This problem can be alleviated by using lager data to modeling the phrase language model. Fortunately, large amounts of monolingual data are recently available on the web.

But using larger data introduces another problem, i.e., n-gram sparseness. Phrase based n-gram has much more vocabulary (i.e., the target phrases observed in a phrase table) than word based n-gram. The increase of the n-gram size is inevitable. The large vocabulary size can cause an efficiency problem that the performance gain from phrase based n-gram becomes too small for the large size of the model. Pruning vocabulary is necessary to reduce the n-gram size. We tested two methods for pruning phrase vocabulary. The first method is a simple singleton pruning, i.e., pruning the singletons when we get the phrase level vocabulary from the phrase table. Another method is to use the phrases which are actually used in a translation. The 'used phrases' are obtained by running the decoder on its training corpus. The used phrases are the phrases that appear in the decoding result. Table 2 shows vocabulary size for each case.The phrase based language model is incorporated in a log-linear model as a feature function analogous to word based language model.

Table 3, 4, and 5 show some experimental results for comparing the results of "with" and "without" a phrase based language model for CRR[1] and ASR[2] input conditions.

[1] Correct recognition result
[2] 1-best Automatic Speech Recognition result

The BLEU scores in the tables are optimized by minimum error rate training. We selectively used one of the three types of phrase language model: full model, model with singleton phrase pruning, and model with used phrases only; and their n-gram order for each test.

Table 2. Vocabulary size of word and phrase LM

	Word	Phrase (full)	Phrase (used)	Phrase (without singleton)
AE	8,171	160,925	43,883	34,574
CE	8,171	121,531	40,651	29,390
CS	10,996	89,005	40,229	24,465

Table 3. Effect of Phrase language model on Arabic-English

	Baseline		With Phrase LM (2-gram)	
	CRR	ASR	CRR	ASR
Dev1	0.4186		0.4202	
Dev2	0.4162		0.4172	
Dev3	0.3965		0.3997	
Dev4	0.1851	0.0848	0.1875	0.0849
Dev5	0.1456	0.1328	0.1446	0.1317
Dev6	0.2973	0.2666	0.2988	0.2641

*Table 4.*Effect of phrase language model on Chinese English task

	Baseline		With Phrase LM (4-gram)	
	CRR	ASR	CRR	ASR
Dev1	0.2247		0.2281	
Dev2	0.2805		0.2865	
Dev3	0.3762	0.2351	0.3770	0.2427
Dev4	0.1212	0.1005	0.1216	0.1033
Dev5	0.1084	0.0919	0.1107	0.0925
Dev6	0.2155		0.2148	

Table 5. Effect of phrase language model on Chinese Spanish task

	Baseline		With Phrase LM (2-gram)	
	CRR	1best	CRR	1best
Dev3	0.2560	0.1691	0.2580	0.1699

The effect of phrase level language model is not statistically significant in the experiments. While trying to find the cause, we noticed that higher order n-grams are also not significant to machine translation quality (see table 6). If the higher order n-gram improves the result, phrase level language model does so (see table 7). The effect of phrase language model is basically analogous to higher order n-grams because it models n-gram dependency upon phrases which consist of one or more words. The phrase n-grams suffer from severe data sparseness as well as higher order n-grams; we have used only 20k given sentences for language modeling.

Table 6. Effect of n-gram for dev3 (BLEU score)

	AE	CE	CS
3gram	0.3965	0.3762	0.2560
4gram	0.3959	0.3803	0.2562
5gram	0.3963	0.3806	0.2533
6gram	0.3965	0.3829	0.2524

Table 7. Comparisons of higher-order ngram with phrase ngram in the official evaluation condition

	Word 3gram	Word 6gram	Word 3gram Phrase 2gram
AE	0.3892	0.4025	0.3940
CE	0.3024	0.2998	0.3039
CS	0.2378	0.2485	0.2570

4. Evaluation Campaign

We have built translation and language models of the SMT systems for IWSLT 2008 evaluation campaign using the only supplied data i.e., 19,972 parallel sentences for each task. The weights of log-linear model are optimized on the entire development set. We merged the development corpora to make a single development corpus. The merged development corpus has seven references. To make the symmetry, we used first seven references for dev1-3 and the first reference was reproduced to make 7th reference for dev6.

We tested our two proposed methods on the development corpus in order to build final system for the evaluation campaign. The results are summarized in table 8. We marked the system applying both of the two methods as primary, and the baseline system built with using Moses without modification as contrastive.

Tokens removed by 'deleting useless' method have been chosen according to deletion test results described in the section 3. We removed tokens of most useless POS for each task, i.e., PUNC for Arabic to English, DEG for Chinese to English, and CS for Chinese to Spanish. N-gram order and pruning type of the phrase based language models are empirically determined to maximize BLEU score on the development corpus for each task. By combining two proposed methods we could improve the MERT results only except Chinese to English ASR 1-best translation.

Table 8. MERT results on development sets (BLEU score)

		Baseline *contrast*	Deleting	PLM	Both *primary*
AE	CRR	0.2700	0.2712	0.2703	0.2718
	ASR	0.1628	0.1657	0.1627	0.1659
CE	CRR	0.1896	0.1922	0.1899	0.1920
	ASR	0.1233	0.1214	0.1239	0.1221
CS	CRR	0.2443	0.2578	0.2551	0.2580
	ASR	0.1677	0.1771	0.1772	0.1782

The official and additional evaluation results for test set are shown in table 8. In the results, the observed changes of

Table 9. Evaluation results

			BLEU	NIST	WER	PER	GTM	METEOR	TER
BTEC_AE case punc	CRR	Primary	0.3878	7.6156	0.4690	0.4198	0.6994	0.6177	41.9660
		Contrast	0.3892	7.5924	0.4662	0.4201	0.6967	0.6167	41.3530
	ASR	Primary	0.2999	6.3244	0.5441	0.4904	0.6306	0.5482	48.6370
		Contrast	0.2973	6.3502	0.5554	0.5011	0.6224	0.5441	49.8150
BTEC_AE no case no punc	CRR	Primary	0.3867	8.1558	0.4742	0.4183	0.6866	0.6172	41.0170
		Contrast	0.3895	8.1078	0.4717	0.4183	0.6843	0.6189	40.4640
	ASR	Primary	0.2929	6.5991	0.5600	0.4976	0.6059	0.5429	49.1490
		Contrast	0.2875	6.6507	0.5754	0.5080	0.6031	0.5407	50.7550
BTEC_CE case punc	CRR	Primary	0.2841	6.3012	0.6179	0.5302	0.6299	0.5104	54.1560
		Contrast	0.3024	6.4593	0.6141	0.5264	0.6308	0.5150	53.6900
	ASR	Primary	0.2624	6.2410	0.6432	0.5546	0.6048	0.4897	57.8100
		Contrast	0.2511	6.0865	0.6557	0.5591	0.5886	0.4851	59.2570
BTEC_CE no case no punc	CRR	Primary	0.3052	7.1788	0.6056	0.4924	0.6591	0.5462	52.9830
		Contrast	0.3212	7.3788	0.6026	0.4896	0.6595	0.5533	52.4600
	ASR	Primary	0.2792	6.9036	0.6415	0.5272	0.6262	0.5199	57.7760
		Contrast	0.2692	6.7507	0.6552	0.5397	0.6129	0.5168	59.2280
BTEC_CS case punc	CRR	Primary	0.2594	5.3343	0.6249	0.5494	0.5728	0.2731	54.2000
		Contrast	0.2378	5.0502	0.6433	0.5752	0.5453	0.2695	56.1500
	ASR	Primary	0.2104	5.4017	0.7335	0.6398	0.5643	0.2658	70.7500
		Contrast	0.2204	5.0648	0.6836	0.6031	0.5297	0.2535	60.6000
BTEC_CS no case no punc	CRR	Primary	0.2537	6.0118	0.6472	0.5445	0.5764	0.2823	56.2900
		Contrast	0.2340	5.9162	0.6553	0.5571	0.5621	0.2827	57.5340
	ASR	Primary	0.1908	5.3856	0.7628	0.6462	0.5650	0.2722	74.7870
		Contrast	0.2150	5.5472	0.6944	0.5941	0.5341	0.2651	61.9350

machine translation quality are not consistent with each other. For correct recognition result translation, our method improved Chinese to Spanish translation but did not so for the others. For 1best ASR output translation, on the other hands, we got completely different results, i.e., our methods improved Arabic to English and Chinese to English translation but Chinese to Spanish. However, table 7 shows that each text translation result with phrase based language model is at least comparable to its baseline. Thus this inconsistency is caused by deleting useless tokens. This means that the uselessness determined by POS tags is very sensitive to input condition. More detailed tag set may be required to alleviate this problem.

The changes made by our methods are very small in Arabic-English task. We had a mistake for Arabic-English task submission. We have found that we deleted the tokens tagged 'NOFUNC' instead of 'PUNC.' The deleted tag has been classified into 'useful' in the test described in the section 3. The performance degradation caused by deleting useful tokens might be canceled with the improvement driven by phrase based language model.

5. Conclusions

The two methods sometimes improved the system and sometimes made it worse. We conclude that the phenomena are a kind of over fitting problem caused by sparse data for phrase based language model; the changes of performance are not depending on language pair and input type.

This IWSLT 2008 evaluation campaign presents a good opportunity to diagnose our methods, especially, phrase based language modeling. Phrase based language model for phrase based machine translation is conceptually sound but have some problematic points. We would continue to make up for the problematic points in phrase based language model for future works.

Acknowledgements

This work was supported by the IT R&D program of MKE/IITA. [2006-S-037, Domain Customized Machine Translation Technology Development for Korean, Chinese, English]

References

[1] 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. and Herbst, E., "Moses: Open Source Toolkit for Statistical Machine Translation", *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007*.

[2] Stolcke, A., "SRILM-an extensible language modeling toolkit," in *Proc. ICSLP*, 2002

[3] Och, F. J., "Minimum error rate training in statistical machine translation," in Proc. 41st Annual Meeting of the Association for Computational Linguistics, pp. 160-167, 2003

[4] Och, F. J., and Ney, H. "A Systematic comparison of various statistical alignment models." *Computational Linguistics*, vol. 29, no. 1, pp. 19-51, 2003.

[5] Papineni, K., Roukos, S., Ward, T., and Zhu, W., "BLEU: a Method for Automatic Evaluation of Machine

Translation," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318, July 2002.

[6] Diab, M., Hacioglu, K., and Jurafsky, D., "Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks," in *proc. HLT-NAACL*, 2004.

[7] Levy, R. and Manning, C., "Is it harder to parse Chinese, or the Chinese Treebank,"? in *proc. 41st Annual Meeting on Association for Computational Linguistics*, pp.439-446, 2003

[8] Koehn, P., Och, F. J., and Marcu, D., "Statistical Phrase-based Translation," in *Proc. HLT-NAACL 2003*, pp. 127-133, 2003.

[9] Och, F. J., and Ney, H., "The Alignment Template Approach to Statistical Machine Translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417-449, 2004.