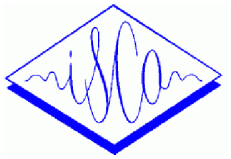
ISCA Archive
http://www.isca-speech.org/archive



International Workshop on Spoken Language Translation (IWSLT) 2006
Keihanna Science City, Kyoto, Japan
November 27-28, 2006

IBM Arabic-to-English Translation for IWSLT 2006

Young-Suk Lee

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
ysuklee@us.ibm.com

Abstract

We present techniques for improving domain-specific translation quality with a relatively high OOV ratio on test data sets. The key idea is to maximize the vocabulary coverage without degrading the translation quality. We maximize vocabulary coverage by segmenting a word into a sequence of morphemes, *prefix*-stem-suffix** and by adding a large amount of out-of-domain training corpora. To preserve the domain-specific meaning of vocabularies occurring in both domain-specific and out-of-domain training corpora, we assign a higher weight to the domain-specific corpus than to the out-of-domain corpora. IBM Arabic-to-English spoken language translation systems using these techniques have demonstrated the best performances in the Open Data Track of the *IWSLT2006 Evaluation Campaign*.

1. Introduction

Creating a large amount of domain-specific parallel corpus by manual translation is very costly and time consuming. And the sparsity of a domain-specific parallel corpus often leads to a very high OOV ratio on unseen data. For example, the OOV ratio of the Arabic-to-English translation development test data for the IWSLT 2006 Evaluation Campaign with respect to the supplied BTEC corpus (Basic Traveler's Expression Corpus), consisting of about 20k sentence pairs, is over 10%.

It has also been noted in previous evaluation campaigns that improving the translation quality of domain-specific evaluation data by adding out-of-domain bilingual corpora is quite challenging. In IWSLT 2004 Chinese-to-English translations, [2], training data for the Small Data track was limited to the supplied BTEC corpus, whereas additional out-of-domain corpora could be used for the Additional Data Track. Tables 1 & 2, illustrate the impact of out-of-domain bilingual corpora on BTEC translation quality.

	BLEU	NIST	GTM	mPER	mWER
CE-S	0.454	8.55	0.720	0.390	0.455
CE-A	0.351	7.39	0.655	0.420	0.496

Table 1. Comparison of the best performing systems in the small (CE-S) and the additional (CE-A) data tracks

Table 1 indicates the performance of the best performing system in the additional (CE-A) data track is consistently worse than that of the best performing system in the small (CE-S) data track.

	BLEU	NIST	GTM	mPER	mWER
S_1-S	0.374	7.74	0.672	0.425	0.488
S_1-A	0.311	5.82	0.632	0.480	0.572
S_2-S	0.349	7.09	0.644	0.430	0.507
S_2-A	0.351	7.39	0.655	0.420	0.496

Table 2. Comparison of two systems (S_1 & S_2) in their small (S) and additional (A) data track submissions

Table 2 shows that the performance of system S_1 is consistently better in the small data track (S_1-S) than in the additional data track (S_1-A). For system S_2, its performance in the additional data track (S_2-A) is only marginally better than in the small data track (S_2-S).

Given the sparsity of a domain-specific parallel corpus in general and the difficulties of improving domain-specific translation quality by adding out-of-domain corpora which exist in a large amount for many language pairs, e.g. Arabic−English, Chinese−English, we present techniques to improve domain-specific translation quality by maximizing vocabulary coverage, which have proven effective for our Arabic-to-English translation systems. Maximization of vocabulary coverage is achieved by (i) segmentation of a word into a sequence of morphemes, *prefix** − *stem* − *suffix** [1] [5], and morphological analysis [6], and (ii) addition of a large amount of out-of-domain corpora. To overcome the problem of performance degradation by adding out-of-domain corpora, we assign a higher weight to the domain-specific training corpus than to the out-of-domain corpora for translation model training. This enables the system to choose the domain-specific

[1] * denotes 0 or more morphemes.

meaning of words/phrases if they occur both in the domain-specific and out-of-domain corpora.

In Section 2, we discuss our baseline phrase translation system. In Section 3, we present the techniques and experimental results. In Section 4, we discuss our system performances in the IWSLT 2006 Evaluation Campaign. In Section 5, we summarize this paper and discuss future work. Throughout this paper, all experiments are carried out on Arabic-to-English translations, and domain-specific corpus refers to the supplied BTEC corpus consisting of about 20k sentence pairs.

2. Baseline Phrase Translation System

IBM Spoken Language Translation systems are based on phrase translation models [4, 11], and DP-based phrase decoder [14]. The baseline system we used in the current evaluation campaign is detailed in [9]. We briefly discuss the key features of block acquisition and decoding. $\bar{e}$ denotes the target phrase and $\bar{f}$ denotes the source phrase.

2.1. Block Acquisition

Blocks are obtained via word alignment and block selection algorithms. We word-align a parallel corpus bi-directionally, using HMM alignments [15]: one from a source word position to a target word position, (A_1: $f \rightarrow e$) and the other from a target word position to a source word position (A_2: $e \rightarrow f$). We define precision (A_P) and recall (A_R) oriented alignments, as in (1):

$$(1) \quad A_P = A_1 \cap A_2 \qquad A_R = A_1 \cup A_2$$

A_P is the intersection of A_1 and A_2, a high precision alignment. A_R is the union of A_1 and A_2, a high recall alignment. Starting from a high precision word alignment A_P, we obtain blocks according to the projection and extension algorithms [8, 13]. We filter out blocks containing non-contiguous source word sequence. For the current evaluation, we also filtered out blocks with a highly improbable source and target phrase length ratio [7].

2.2. Decoding

The phrase decoder utilizes 10 distinct scoring functions multiplied by their respective weight:[2]

- Direct phrase translation model cost
- Source-channel phrase translation model cost
- Block unigram model cost
- IBM Model 1 cost applied in both directions
- Word trigram language model costs
  - For the first word of a target phrase
  - Subsequent words of a target phrase
- Word & block count penalty
- Outbound and inbound distortion model costs

Direct phrase translation model probabilities are computed according to (2).

$$(2) \quad p(\bar{e} \mid \bar{f}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{e}'} count(\bar{e}', \bar{f})}$$

Source-channel model probabilities are computed according to (3).

$$(3) \quad p(\bar{f} \mid \bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}'} count(\bar{f}', \bar{e})}$$

Unigram probability of a block ($b = \bar{e}, \bar{f}$), is computed according to (4):

$$(4) \quad p(b) = \frac{count(b)}{\sum_{b'} count(b')}$$

IBM Model 1 translation cost is computed according to (5) for each block.

$$(5) \quad \sum_{j=1}^{m} -\log_{10} \max p(f_j \mid e_i), 1 \le i \le n$$

j is the source word position index (m is the number of source words in the source phrase). i is the target word position index (n is the number of target words in the target phrase). If $max\ p(f_j \mid e_i)$ is 0.0, and therefore $-\log_{10} \max p(f_j \mid e_i)$ is infinite, we assign a fixed cost β for f_j, which is empirically determined on the basis of training corpus size and the properties of the given language pair. Model 1 translation probability $p(f_j \mid e_i)$ is approximated by the relative frequency of blocks consisting of one source word and one target word, as in (6).

$$(6) \quad p(f \mid e) = \frac{count(f, e)}{\sum_{f'} count(f', e)}$$

Trigram language model is used for the target phrase ($\bar{e}$). We assign different weights to the first word e_1 and the remaining word e_i ($1 < i \le n$, n: number of

[2] We use manually tuned weights since manually tuned weights lead to better system performances than automatically tuned weights.

words in the target phrase). Word and block count penalties are applied to avoid the general tendency to choose the shortest translation output and the longest matching source phrase, [16]. Word level distortion models [1], along with the skip operation, are used to further improve the word order accuracy.

3. Technique Proposal

We present the techniques which simultaneously maximize the vocabulary coverage and improve the translation quality of test data.

The list of parallel training corpora used in all of our experiments is given in Table 3. # AR word denotes the number of punctuation-tokenized Arabic words, # EN word denotes the number of punctuation-tokenized English words and Size denotes the number of sentence pairs. OOD Total indicates the statistics for the out-of-domain corpora, i.e. excluding the BTEC corpus.

Source	# AR word	# EN word	Size
BTEC	**159,213**	**189,239**	**20,000**
LDC2003T18	26,146	33,869	1,043
LDC2003E05	103,717	129,181	4,235
LDC2003E09	123,505	150,865	5,003
LDC2004E07	520,971	681,613	20,358
LDC2004E11	227,792	310,079	8,576
LDC2004E08	1,771,893	2,207,934	52,042
LDC2005E46	616,879	819,354	24,874
LDC2001T55	70,183	80,354	2,346
FBIS	86,614	117,420	2,624
OOD Total	**3,547,700**	**4,530,669**	**121,119**

Table 3. Size of the parallel training corpora

3.1. Maximization of Vocabulary Coverage

We increase the vocabulary coverage by adding out-of-domain training corpora, segmentation of a word into a sequence of *prefix*-stem-suffix**, and morphological analysis.

OOV ratio of the IWSLT2005 evaluation data set (EVAL05) and the IWSLT2006 development test data set (DEV06) on the BTEC corpus without Arabic word segmentation (baseline), with Arabic word segmentation (segment) and Arabic morphological analysis (analysis), is given in Table 4.

	EVAL05	DEV06
baseline	4.96% (157/3164)	10.27% (489/4763)
segment	1.18% (56/4747)	2.54% (195/7671)
analysis	2.04% (80/3914)	4.41% (271/6147)

Table 4. OOV ratio on the BTEC training corpus

Vocabulary count for the EVAL05 includes human punctuations and that for DEV06 includes automatic machine punctuations. OOV ratio of EVAL05 and DEV06 with respect to the BTEC corpus plus the out-of-domain training corpora is given in Table 5.

	EVAL05	DEV06
baseline	2.18% (69/3164)	4.66% (222/4763)
segment	0.51% (24/4747)	1.19% (91/7671)
analysis	0.79% (31/3914)	1.82% (112/6147)

Table 5. OOV ratio of the development test data sets on the BTEC and the out-of-domain corpora

3.1.1. Arabic Word Segmentation

Arabic has very rich inflections including person, number, gender, case, etc. Furthermore, an Arabic word (demarcated by a white space) often corresponds to more than one English word, decomposable into several morphemes in the sequence of *prefix*-stem-suffix**. For the Arabic-English parallel sentence (7), an ideal word alignment may be represented, as in (8), where each of the two Arabic words (written in Buckwalter transliteration) *Aryd* and *AzAlthA* is aligned to two English words:

(7) a. lA Aryd AzAlthA
b. I don't want it extracted

(8) lA <==> don't
Aryd <==> I want
AzAlthA <==> it extracted

We increase the vocabulary coverage by segmenting a word into morphemes, as in (9), where # denotes a prefix, and + a suffix:

(9) a. Aryd → A# ryd
b. AzAlthA → AzAl +t +hA

We use a language model based Arabic word segmenter, [5, 10], which segments an input text into a sequence of morphemes using the language model parameters estimated from word segmented Arabic training corpus. A sample un-segmented Arabic text and its segmented output are shown in Figure 1. Multiple prefixes and suffixes per word are underlined.

wsyHl sA}q AltjArb fy jAgwAr AlbrAzyly lwsyAnw bwrty mkAn AyrfAyn fy AlsbAq gdA AlAHd Al*y sykwn Awly xTwAth fy EAlm sbAqAt Alfwrmwlā
w# s# y# Hl sA}q Al# tjArb fy jAgwAr Al# brAzyly lwsyAnw bwrty mkAn AyrfAyn fy Al# sbAq gdA Al# AHd Al*y **s# y#** kwn Awly xTw **+At +h** fy EAlm sbAq +At Alfwrmwlā

Figure 1. Arabic text before (top) and after (bottom) word segmentation

3.1.2. Morphological Analysis

Since we do not segment inflectional suffixes in English, there are some Arabic suffixes which get incorrectly aligned. To accomplish better word-to-word correspondences between the source and the target languages, we merge some Arabic prefixes/suffixes to their stems or delete them, morphological analysis. The algorithm for automatically determining which morphemes to be merged or deleted is detailed in [6].

Arabic text after applying morphological analysis to word segmented Arabic corpus is given in Figure 2. Merged prefixes and suffixes are underlined and deleted ones are denoted by **ø**.

w# s# **y**Hl sA}q **ø** tjArb fy jAgwAr Al# brAzyly lwsyAnw bwrty mkAn AyrfAyn fy Al# sbAq gdA **ø** AHd Al*y s# **y**kwn Awly xTw**At** +h fy EAlm sbAq**At** Alfwrmwlā

Figure 2. Arabic morphological analysis

Table 6 shows the vocabulary size of the BTEC and BTEC plus the additional out-of-domain corpora, without Arabic word segmentation (baseline), with Arabic word segmentation (segment) and Arabic morphological analysis (analysis).

corpora	BTEC		BTEC + additional	
language	Arabic	English	Arabic	English
baseline	17,278	7,159	136,577	57,597
segment	7,952	7,159	40,264	57,597
analysis	10,717	7,159	59,808	57,597

Table 6. Vocabulary size of the BTEC and BTEC plus out-of-domain training corpora

3.1.3. System Performances

Translation quality of EVAL05 & DEV06 translated by the systems trained on the BTEC corpus are shown in Table 7 in BLEU [12].[3]

	EVAL05		DEV06	
	BLEU	95% conf	BLEU	95% conf
baseline	0.5622	+/−0.0323	0.2474	+/−0.0203
segment	0.6043	+/−0.0309	0.2886	+/−0.0213
analysis	0.5880	+/−0.0312	0.2973	+/−0.0216

Table 7. Translation quality of the systems trained on the BTEC corpus

[3] We use the BLEU script implemented by K. Papineni which computes the brevity penalty on the basis of the closest matching reference translation in length, as opposed to NIST-implemented mteval-11b.pl which computes the brevity penalty on the basis of the shortest matching reference translation.

Translation quality of the systems trained on the BTEC corpus plus the out-of-domain corpora are shown in Table 8.

	EVAL05		DEV06	
	BLEU	95% conf	BLEU	95% conf
baseline	0.5619	+/−0.0291	0.2544	+/−0.0225
segment	0.5705	+/−0.0299	0.2968	+/−0.0205
analysis	0.5720	+/−0.0311	0.2943	+/−0.0214

Table 8. Translation quality of the systems trained on the BTEC plus out-of-domain parallel corpora

Translation quality of EVAL05 are measured with 16 reference translations, and DEV06 with 7 reference translations. Both data sets are scored with upper/lower case distinctions and punctuations, computing up to 4-gram precisions. 95% conf indicates the BLEU scores to be added (+) and subtracted (−) to be statistically significant at 95% confidence interval.

Tables 7 & 8 indicate that Arabic word segmentation and morphological analysis improve the translation quality significantly regardless of whether the system is trained on the BTEC corpus only, or BTEC corpus plus out-of-domain training corpora. Comparison between Table 7 and Table 8, however, indicates that addition of out-of-domain corpora hurts the performance of EVAL05 which has a relatively low OOV ratio on the BTEC corpus, while the additional corpora improve the performance of DEV06 somewhat which has a relatively high OOV ratio on the BTEC corpus, cf. Table 4.

3.2. Domain-Specific Meaning Preservation

We assess that the translation quality degradation of EVAL05 after adding the out-of-domain training corpora is to be ascribed to the following: Potential performance improvement due to the increased vocabulary coverage is overridden by the performance degradation caused by incorrect meaning selection of words/phrases which occur both in the domain-specific and out-of-domain training corpora. We would like to remind the reader that the size of the out-of-domain training data is at least 22 times bigger than that of the BTEC corpus in word counts, cf. Table 3.

To overcome this problem, we train the translation models by assigning a higher weight to the BTEC corpus than to the out-of-domain corpora. Figure 3 and Figure 4 show the system performances as a function of the weights assigned to the BTEC and the out-of-domain training corpora. The thinner solid line indicates the baseline system performances, and the thicker solid line indicates performances of the systems trained on word segmented Arabic corpus. The dotted line indicates performances of the systems trained on the morphologically analyzed Arabic corpus.

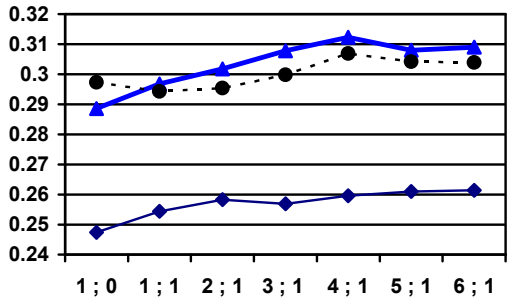


Figure 3. System performances on DEV06 as a function of weights assigned to the BTEC vs. out-of-domain training corpora

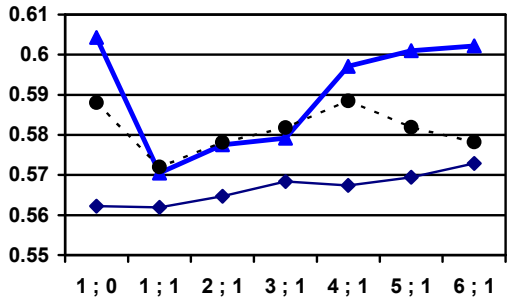


Figure 4. System performances on EVAL05 as a function of weights assigned to BTEC vs. out-of-domain training corpora

Y-axis indicates cased BLEU scores, x-axis, the ratio between the BTEC corpus and the out-of-domain corpora. 1;0 denotes that only the BTEC corpus is used. 2;1 denotes that weight 2 is assigned to the BTEC corpus and 1 to the out-of-domain corpora, etc. The weight ratio change in the training corpus is reflected in the following decoder scoring functions: direct phrase translation model, source channel phrase translation model, block unigram model, and IBM Model 1 costs.

Figure 3 shows that the translation quality of DEV06 steadily improves up to the weight ratio reaches 4;1. For the systems trained on word segmented Arabic corpus, performance improvement is 0.0237 BLEU points (0.2886 → 0.3123), for systems trained on morphologically analyzed Arabic, improvement is 0.0096 (0.2973 → 0.3069). Performance improvement of the baseline system is 0.0136 (0.2474 → 0.2614). Figure 4 indicates that translation quality improvement of EVAL05 is not as clear-cut as that of DEV06. Translation quality degrades when the systems are trained on the BTEC corpus plus out-of-domain corpora with an equal weight (1;1), compared with the system trained on the BTEC corpus (1;0). However, translation quality improves as we assign a higher weight to the BTEC corpus. Performances of the systems trained on un-segmented Arabic improve by 0.0107 BLEU points (0.5622 → 0.5729) as the weight ratio reaches 6;1. Translation quality of the systems trained on word segmented (0.6043 → 0.6022) and morphologically analyzed (0.5880 → 0.5885) Arabic remain virtually the same.

3.3. System Combination

Once we develop various translation systems whose translation lexicons vary according to (i) the domain-specificity of the parallel corpus, and (ii) Arabic corpus processing − un-segmented, word segmented, morphologically analyzed − we apply the system combination algorithm we developed for the IWSLT 2005 Evaluation Campaign [8], cf. [17].

The key aspect of the algorithm is to choose the translation output of the system with the lowest translation cost (i.e. the best translation output) among various system outputs for each translation segment. We pre-determine the system which generally results in the highest translation quality measured by BLEU. We call the system producing the highest translation quality *h-sys*, and the systems producing lower translation quality, $l\text{-}sys_1$, … $l\text{-}sys_n$. If the translation cost of a lower-performing system $l\text{-}sys_n$ is lower than that of the highest-performing system *h-sys*, we choose the translation output of $l\text{-}sys_n$.

One of the most effective system combinations for DEV06 and EVAL05 is shown in Table 9 and Table 10, respectively.

	Training corpora weights	Arabic processing	BLEU
h-sys	4 BTEC; 1 out-of-domain	word segmentation	0.3123
$l\text{-}sys_1$	4 BTEC; 1 out-of-domain	morpho analysis	0.3069
$l\text{-}sys_2$	1 BTEC; 0 out-of-domain	morpho anaysis	0.2973
System combination: h-sys+l-sys$_1$+l-sys$_2$			**0.3245**

Table 9. Effective system combination for DEV06

	Training corpora	Arabic processing	BLEU
h-sys	1 BTEC; 0 out-of-domain	word segmentation	0.6043
$l\text{-}sys_1$	4 BTEC; 1 out-of-domain	word segmentation	0.5971
$l\text{-}sys_2$	4 BTEC; 1 out-of-domain	morph analysis	0.5885
System combination: h-sys+l-sys$_1$+l-sys$_2$			**0.6200**

Table 10. Effective system combination for EVAL05

Correct Recognition Result					
Scoring Method	BLEU4	NIST	METEOR	WER	PER
Official	***0.2549***	6.3769	0.5316	0.5668	0.4825
Additional	***0.2773***	7.1681	0.5314	0.5593	0.4480
ASR Output					
Scoring Method	BLEU4	NIST	METEOR	WER	PER
Official	***0.2274***	5.8466	0.4845	0.6049	0.5198
Additional	***0.2428***	6.4867	0.4842	0.6035	0.4958

Table 11. IBM Arabic-to-English Spoken Language Translation System Performance in IWSLT 2006 Open Data Track

4. IWSLT2006 Evaluation Campaign

We have participated in the Arabic-to-English Open Data Track. Automatic scoring results are shown in Table 11. Official submissions were scored with punctuations and case information while additional submissions were scored without them. IBM systems have demonstrated the best performances across all evaluation conditions: (i) Correct Recognition Result + Official, (ii) Correct Recognition Result + Additional, (iii) ASR output + Official, (iv) ASR output + Additional.

4.1. Pre-processing

Evaluation data were distributed without punctuations, and we restored them in the pre-processing stage, using the LM-based punctuation restorer described in [9]. All of the Arabic pre-processing steps are listed below:

1) Spelling normalization of alif variants
2) Punctuation restoration
3) Rule-based number classing
4) Word segmentation[4]
5) Part-of-speech tagging
6) Morphological analysis

We apply (1) to (3) to derive the translation lexicon of un-segmented Arabic, (1) through (4) to derive the translation lexicon of word-segmented Arabic, and (1) through (6) to derive the lexicon of morphologically analyzed Arabic. Part-of-speech tagging is needed for morphological analysis. We lowercase all English words.

4.2. Decoding

Pre-processed input segments are decoded by the phrase decoder, cf. Section 2.

For trigram language model training we use both the BTEC and out-of-domain training corpora summarized in Table 12.

BTEC	English Gigaword
~380k words from ~190k AE supplied/IWSLT06 ~190k JE supplied/IWSLT04	~2.5 billion words from LDC2005T12

Table 12. LM training corpus statistics

Analogous to the technique we use for the translation model training, we assign a higher weight to the LM derived from the BTEC corpus (0.7) than the one derived from the out-of-domain corpora (0.3).

We have tuned the decoder parameters on the Correct Recognition Results of the DEV06, and used the same parameters for the Correct Recognition Result and the ASR output in the official evaluation.

4.3. System Combination and Post-processing

Translation model training corpora weights and the Arabic processing of the three systems we used for the system combination are shown in Table 13.

	training corpora weights	Arabic processing
S_1	4 BTEC;1 out-of-domain	segmentation
S_2	4 BTEC;1 out-of-domain	morph analysis
S_3	1 BTEC;0 out-of-domain	morph analysis

Table 13. Training corpora and Arabic corpus processing of the systems deployed in IWSLT 2006

Table 14 shows the OOV ratio of the Correct Recognition Result on S_1, S_2, S_3. It also shows the corresponding statistics on two additional systems: the baseline system S_0 trained on the BTEC corpus without Arabic word segmentation and S_0' trained on BTEC corpus with Arabic word segmentation.

Token Count includes automatically inserted punctuations. Without punctuations, the OOV ratio on the S_0 (baseline) is 13.09% (609/4654), much higher than 11.65% with punctuations.

[4] We apply reordering between the future tense prefix *s#* and other verbal prefixes *n#, t#, y#* to the word segmented evaluation data. This reordering has minimally hurt the performance in our submission runs (from 0.2576 to 0.2549 for the Correct Recognition Result & from 0.2288 to 0.2274 for the ASR output).

	OOV Count / Ratio	Token Count
S_0 (baseline)	609 / 11.65 %	5229
S_0'	196 / 2.33 %	8420
$\boldsymbol{S_1}$	82 / 0.97 %	8420
$\boldsymbol{S_2}$	101 / 1.48 %	6805
$\boldsymbol{S_3}$	339 / 4.98 %	6805

Table 14. OOV Ratio of Correct Recognition Result

Each component and the combined system performances in the official submissions are given in Table 15. Count indicates the number of segments chosen from each system for $\boldsymbol{S_1}$, $\boldsymbol{S_2}$ and $\boldsymbol{S_3}$, and the total number of segments for $\boldsymbol{S_1}$+$\boldsymbol{S_2}$+$\boldsymbol{S_3}$. BLEU scores of the baseline system S_0 and S_0' are given for comparisons:

	Correct Recognition		ASR Output	
	BLEU	Count	BLEU	Count
S_0	0.2224	N/A	0.2102	N/A
S_0'	0.2352	N/A	0.2117	N/A
$\boldsymbol{S_1}$	0.2533	309	0.2243	289
$\boldsymbol{S_2}$	0.2442	115	0.2175	121
$\boldsymbol{S_3}$	0.2349	76	0.2148	90
$\boldsymbol{S_1}$+$\boldsymbol{S_2}$+$\boldsymbol{S_3}$	**0.2549** +/−0.0182	500	**0.2274** +/−0.0177	500

Table 15. Component and combined system performances in the IWSLT 2006 official submissions

The performance improvements from $\boldsymbol{S_3}$ to $\boldsymbol{S_2}$ (0.2349 → 0.2442 & 0.2148 → 0.2175) is due to the addition of out-of-domain corpora with the BTEC vs. out-of-domain corpora weight ratio of 4;1. The difference between $\boldsymbol{S_2}$ and $\boldsymbol{S_1}$ (0.2442 vs. 0.2533 & 0.2175 vs. 0.2243) indicates word segmentation is more effective than the morphological analysis, given the specified BTEC and out-of-domain training corpus combination. The gap between $\boldsymbol{S_1}$+$\boldsymbol{S_2}$+$\boldsymbol{S_3}$ and $\boldsymbol{S_1}$ (0.2549 − 0.2533 & 0.2274 − 0.2243) is the performance gain due to system combination.

Post-processing (i) restores upper/lower case distinction using word trigram language models, (ii) merges contracted words *'m*, *'ll*, *'re*, *'ve*, *'s*, *'d* into the preceding words, as in *we 're* → *we're*, in the English translation output.

5. Summary and Future Work

We have presented techniques for improving domain-specific translation quality. We have used Arabic word segmentation and morphological analysis to increase the vocabulary coverage on unseen data. We have also added a large amount of out-of-domain training corpora (more than 6 times bigger than the BTEC corpus in terms of sentence pair counts). To avoid translation quality degradation resulting from adding a large out-of-domain corpus, we have assigned a higher weight to the BTEC corpus than to the out-of-domain training corpora. IBM Arabic-to-English spoken language translation system using these techniques demonstrated the best performances in all evaluation conditions of the Open Data Track.

Increase in vocabulary coverage by segmenting a word into morphemes should be applicable to other languages with rich morphology such as Korean. It should be particularly effective if there is not enough parallel training corpus on the same domain or when there is a genre mismatch between the training corpus and the evaluation corpus. Preservation of domain-specific meaning of words/phrases occurring in both domain-specific and out-of-domain corpora, by assigning a higher weight to the domain-specific corpus, should be applicable to any genre such as newswire, broadcast news, etc. However, it needs to be further investigated on how to automatically determine the weights to be assigned to the domain-specific and out-of-domain corpora.

Acknowledgements

This work has been funded in part by the European Commission under the project TC-STAR (Technology and Corpora) FP6-506738. We would like to thank the organizers of the IWSLT 2006 Evaluation Campaign and an anonymous reviewer for helpful comments.

References

[1] Y. Al-Onaizan, P. Kishore. "Distortion Models for Statistical Machine Translation", *Proceedings of COLING/ACL 2006*. July 17−21, 2006, Sydney, Australia.

[2] Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, Jun'ich Tsuhii, "Overview of the IWSLT04 Evaluation Campaign", *IWSLT 2004 Proceedings*, pages 1−12, September 30−October 1, 2004, Kyoto, Japan.

[3] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer. "The mathematics of statistical machine translation: parameter estimation", *Computational Linguistics*, 19(2):263−311, 1993.

[4] P. Koehn, F. J. Och, and D. Marcu. "Statistical phrase-based translation", *Proceedings of HLT−NAACL 2003,* pages 48−54, 2003.

[5] Y-S. Lee, K. Papineni, S. Roukos, O. Emam, and H. Hassan. "Language model based Arabic word segmentation", *Proceedings of the 41st Annual Meeting of ACL 2003*, pages 399−406, 2003.

[6] Y-S. Lee. "Morphological analysis for statistical machine translation", *Proceedings of HLT−NAACL 2004: Companion Volume*, pages 57−60, 2004.

[7] Young-Suk Lee and Salim Roukos, "IBM Spoken Language Translation System Evaluation", *IWSLT 2004 Proceedings*, pages 39−46, September 30—October 1, 2004, Kyoto, Japan.

[8] Y-S. Lee. "IBM Statistical Machine Translation for Spoken Languages", *IWSLT 2005 Proceedings*. pages 86−93, October 24−25, 2005, Pittsburgh, USA.

[9] Y-S. Lee, S. Roukos, Y. Al-Onaizan, K. Papineni. "IBM Spoken Language Translation System". *Proceedings of TC-STAR workshop on Speech-to-Speech Translation*, pages 13−18, June 19−21, 2006, Barcelona, Spain.

[10] X. Luo and S. Roukos. "An iterative algorithm to build Chinese language models", *Proceedings of the Annual Meeting of ACL 1996*, pages 139−143, 1996.

[11] F. J. Och, C. Tillmann, and H. Ney. "Improved alignment models for statistical machine translation", *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20−28, 1999.

[12] K. Papineni, S. Roukos, T. Ward, and W. Zhu. "Bleu: A method for automatic evaluation of machine translation", *Proceedings of the 40th Annual Meeting of ACL 2002*, pages 311−318, 2002.

[13] C. Tillmann. "A projection extension algorithm for statistical machine translation", *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 1−8, 2003.

[14] C. Tillmann and H. Ney. "Word reordering and a DP beam search algorithm for statistical machine translation", *Computational Linguistics*, 29(1):97−133.

[15] S. Vogel, H. Ney, and C. Tillmann. "HMM-based word alignment in statistical translation", *Proceedings of COLING−96*, pages 836−841, 1996.

[16] R. Zens and H. Ney. "Improvements in phrase-based statistical machine translation", *Proceedings of HLT-NAACL 2004*, pages 257−264, 2004.

[17] M. Paul, T. Doi, Y. Hwang, K. Imamura, H. Okuma and E. Sumita. "Nobody is Perfect: ATR's Hybrid Approach to Spoken Language Translation", *IWSLT 2005 Proceedings*. pages 86−93, October 24−25, 2005, Pittsburgh, USA.