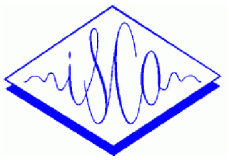
ISCA Archive
http://www.isca-speech.org/archive



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

The UKA/CMU Statistical Machine Translation System for IWSLT 2006

Matthias Eck[†], *Ian Lane*[*], *Nguyen Bach*[*], *Sanjika Hewavitharana*[*], *Muntsin Kolss*[†], *Bing Zhao*[*], *Almut Silja Hildebrand*[†], *Stephan Vogel*[*] *and Alex Waibel*[†*]

InterACT Research Laboratories:
[†]University of Karlsruhe, Karlsruhe, Germany
[*]Carnegie Mellon University, Pittsburgh, USA
http://interact.ira.uka.de/

Abstract

This paper describes the UKA/CMU statistical machine translation system used in the IWSLT 2006 evaluation campaign. The system is based on phrase-to-phrase translations extracted from a bilingual corpus. We compare two different phrase alignment techniques both based on word alignment probabilities. The system was used for all language pairs and data conditions in the evaluation campaign translating both the ASR output (as 1best) and the correct recognition results.

1. Introduction

The UKA/CMU statistical machine translation system that was used for the IWSLT 2006 evaluation campaign is based on phrase to phrase translations. A phrase-to-phrase translation system uses phrases as the general building blocks of the final translation. This generally leads to better translation performance than purely word based translation systems. The main reason is that the phrases are able to preserve local context information thus leading to better lexical choice. In addition a phrase translation can already automatically perform a local re-ordering within the translated phrase. Section 3 presents a detailed look at the UKA/CMU translation system and gives an overview over phrase alignment, language models and decoder architecture. Section 4 presents the results of the experiments done during and after IWSLT 2006. We compare our official submission results with contrastive runs using different settings and conditions.

The IWSLT 2006 evaluation campaign allowed the use of freely available data for the Open data track submissions and also the additional use of proprietary data for the C-STAR data track submissions. The overview will also list additional data that was added for each condition.

The UKA/CMU statistical translation system was used for every language pair for both data conditions. We did not use any specific techniques for ASR output translation so just the 1-best ASR output was translated and compared to the translation of the correct recognition result.

2. InterACT Partner cooperation

The work presented here, is the collaborative effort of researchers at our UKA and CMU laboratories, who combined research systems to test and evaluate approaches on multiple languages. In this manner, we have been able to evaluate on more conditions than any other site and to compare across languages. As shall be seen, however, performance still appears to depend more on the maturity of the effort in any given language than on inherent difficulties of that language. The UKA contributions have focused on applying a Chinese → English translation system that is being developed for the TC-STAR project. Also a first trial with Italian was carried out to further integration and collaboration with other TC-STAR systems. CMU focused on Arabic, a language that is studied under projects GALE and Transtac and also applied the system to Japanese.

The translation system that was used is basically the same in both places. It has to be seen as a combined effort of a close-knit research group [1],[2].

3. Translation System

3.1. Phrase Alignment

3.1.1. Overview

We used two phrase alignment methods for IWSLT 2006, namely *PESA* and *LogLin*. The LogLin phrase extraction method is computationally intensive so it was not used for all submitted systems. More extensive experiments with contrastive systems showed that LogLin generally improves the results compared to the PESA phrase extraction method.

3.1.2. PESA: Phrase pair extraction as sentence splitting [3]

The PESA phrase extraction method is based on the well known IBM-1 word alignment model [4]. The IBM-1 model assigns a probability to all possible word alignments of respective sentences in the training data.

Assuming a sentence in the bilingual corpus contains a phrase from a source sentence $e_{i_1}^{i_2} = e_{i_1}...e_{i_2}$ we are inter-

ested in the sequence of words $f_{j_1}^{j_2} = f_{j_1}...f_{j_2}$ from the respective target sentence that is the optimal translation for this source phrase.

We can now estimate the quality of a translation candidate by using the IBM-1 word alignment probabilities between the source and target phrases.

If the candidate is actually a good translation of the source phrase we expect higher IBM-1 probabilities between the words in the phrases than if the translation candidate was incorrect.

If we assume that $f_{j_1}^{j_2}$ is the optimal translation for $e_{i_1}^{i_2}$ in this sentence pair we can analogously argue that the words from the sentence pair that are not in these phrases must also be translations of each other.

This means the optimal translation for the (non-contiguous) source phrase $e_1...e_{i_1-1} \quad e_{i_2+1}...e_I$ is $f_1...f_{j_1-1} \quad f_{j_2+1}...f_J$ and we also expect high probabilities between the words in these two phrases.

Overall the constrained probability for this sentence split can be calculated as:

$$p_{j_1,j_2}(e|f) = \prod_{i=1}^{i_1-1} \sum_{j \notin (j_1...j_2)} p(e_i|f_j) * \prod_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} p(e_i|f_j) \prod_{i=i_2+1}^{I} \sum_{j \notin (j_1...j_2)} p(e_i|f_j)$$

If we optimize over the target side boundaries j_1 and j_2 we can determine the optimal sentence splitting and the best translation candidate.

The same ideas can be applied if we use the IBM-1 probabilities for the reverse direction thus calculating $p_{j_1,j_2}(f|e)$ and we interpolate the two phrase alignment probabilities to get the optimal translation candidate.

In the actual system we not only use the top translation candidate but all candidates to a certain threshold. This covers translation alternatives and leaves the final decision to other models, mainly the language model.

3.1.3. LogLin: Phrase pair extraction with Log-Linear Features [5],[6]

Another phrase extraction method applied was *LogLin*. LogLin formulates the *phrase-extraction* problem as a *local search*, guided by a simple *heuristic function*.

Given the source n-gram e_i^{i+k} (span from position *i* to position *i+k*, with a length of *k+1*), the local search starts by first localizing the projected center of the target phrase, and then it searches the best scored width, corresponding to the left- and right- boundaries (*j*, *j+l*) of the target phrase: f_j^{j+l} .

Previously the heuristic function included phrase-level fertility score, a simple IBM-1 lexicon score, and a phrase-level position distortion score.

In this evaluation, we used an extended heuristic function: *a log-linear model*, in which thirteen feature functions were computed to predict the qualities of a phrase alignment [6]. The log-linear model allows the utilization of overlapping feature functions.

The thirteen feature functions used are summarized as follows. There are *four* feature functions of phrase-level fertility score: $P(l+1|e_i^{i+k})$ and $P(J-l-1|e_{i' \notin [i,i+k]})$, which are the probabilities to generate length of *l+1* using the source candidate phrase e_i^{i+k} and the remaining length of (*J-l-1*) using the remaining words in the source sentence: $e_{i' \notin [i,i+k]}$. These probabilities are computed using the word fertility $P(\phi|e_i)$ via dynamic programming. Similarly, the probabilities are computed in the other direction as $P(k+1|f_j^{j+l})$ and $P(I-k-1|f_{j' \notin [j,j+l]})$.

Four feature functions compute the IBM Model-1 scores for each candidate phrase-pair: $P(f_j^{j+l}|e_i^{i+k})$ and $P(e_i^{i+k}|f_j^{j+l})$. The remaining parts of the sentence pair (e, f) excluding the candidate phrase-pair is also modeled with $P(f_{j' \notin [j,j+l]}|e_{i' \notin [i,i+k]})$ and $P(e_{i' \notin [i,i+k]}|f_{j' \notin [j,j+l]})$.

Another *four* scores are aimed at bracketing the sentence-pair along the diagonal and the inverse-diagonal using the IBM Model-1 lexicon probabilities of *p(e | f)* and *p(f | e)*.

The final feature function is the average number of word alignment links per source word in the candidate phrase-pair. We assume that each aligned phrase-pair should contain at least one word alignment link (for details concerning the features please see [6]).

To learn the weights for each of the feature function, *gold-blocks* from human word alignments were extracted, and a log-linear model was trained to optimize the accuracy at the phrase-level for each sentence pair. Details about the training can also be found in [6]. In this evaluation, we simply copied the learned weights from our previous experiments during the NIST evaluation.

3.2. Language Model

The UKA/CMU-SMT system supports standard n-gram language models. Here we applied a 6-gram language model using a suffix array implementation .

The suffix array language model provides the possibility for histories of arbitrary length but we only used a history of 5 (effectively a 6-gram language model). This language model uses Good-Turing smoothing (see [7]).

3.3. Decoding [8]

After the preparation and training of translation and language models is complete all models are used in a decoder to translate the actual source sentences.

The UKA/CMU-SMT decoder uses a 2 stage process that first builds a translation lattice and then searches for the best path through the lattice.

The translation lattice is built by using all available translation pairs from the translation models for the given source sentence and inserting them into a lattice. These translation pairs consist of words or phrases on the source side that cover

a part of the source sentence. The decoder inserts an additional edge for each phrase pair and attaches the target side of the translation pair and translations scores to the edge.

The translation lattice will now contain a large number of possible paths that cover each source word exactly once (a combination of partial translation of words or phrases). These translation hypotheses will greatly vary in quality and the decoder uses the different knowledge sources and scores to find the best path possible translation hypothesis.

This step also allows for limited reordering within the found translation hypotheses.

The features, which are used to score each phrase-pair for decoding process, are different from the features used for phrase-extraction in section 3.1.3, which are aimed to *bracket* the *sentence-pairs* given a phrase-pair block and these features are specific/relative to each sentence-pair.

In our decoding experiments for IWSLT06, we used the following scores[1]: relative frequencies in both directions, phrase-level fertility scores in both directions (via DP as in section 3.1.3), The IBM-1 Viterbi scores in both directions, un-normalized IBM-1 lexicon scores in both directions $P(f_j^{j+l}|e_i^{i+k}) = \prod_{j' \in [j,j+l]} \prod_{i' \in [i,i+k]} P(f_{j'}|e_{i'})$ (favoring long phrase-pairs, see [2]), the phrase-level normalized frequency, and the normalized number of alignment links within the phrase-pair (as in section 3.1.3).

4. Translation Results

This section gives an overview of all official and contrastive results achieved by the UKA/CMU translation systems.

All results will only be given in BLEU([9]) and NIST([10]) scores according to the official evaluation specifications (mixed case with punctuation marks). For other scores for the submitted systems please refer to the official scoring publication of IWSLT 2006.

Submissions were done for all language pairs in the Open and C-STAR data conditions. We always translated ASR output (as 1-best) and as a comparison the correct recognition results (CRR).

The development set numbers always refer to the development set that was also provided for the evaluation campaign for IWSLT 2006.

4.1. Data Conditions

For each language pair 20,000 or 40,000 sentences from the BTEC corpus [11] were provided to the participants (Supplied data). For the Open data track it was possible to use any freely available data in addition to this.

As a C-STAR member UKA/CMU also has access to the whole BTEC training corpus for each language (different sizes per language). The actual C-STAR data track allowed the use of this data and also the use of additional free and proprietary data (Full BTEC + any data).

[1] which are not necessarily probabilities

Overall we identified four different data situations that could be investigated.

Data situation	**Explanation**
Supplied	Supplied data only
Supplied + free data	Supplied data and freely available data (*Open data track* in Evaluation campaign)
Extended BTEC	Full BTEC corpus
Full BTEC + any data	Full BTEC corpus and any other data (*C-STAR data track* in Evaluation campaign)

Table 1: *Data sets and conditions*

4.2. Input Conditions

The test data for all language pairs was provided as speech input, ASR output in the form of n-best lists and lattices and as the "correct" recognition results (manual transcription).

We only participated in the ASR output translation and it was always required to also generate a translation for the correct recognition results (using the same system as was used to translate the ASR output) in order to analyze the impact of recognition errors on the translation performance. For the ASR output translation only the 1-best recognition result was used. The testset word accuracy of the 1-best ASR output ranges between 68.11% for Chinese (spontaneous) and 85.14% for Japanese. The word accuracies for Chinese (read), Arabic and Italian are 73.64%, 73.88% and 70.88% respectively. The word accuracy certainly affects the translation performance and we can expect to see lower scores for the Chinese spontaneous speech compared to the read speech.

4.3. Arabic → English

4.3.1. Training Corpora

For Arabic about 20,000 lines of data were provided as Supplied data. We also had an additional 20,000 lines of BTEC data translated to Arabic and added this data for the C-STAR data track submission. Some of the sentences were filtered out due to apparent discrepancies so the actual numbers of lines are a little lower than 20,000/40,000 (see Table 2).

Data Condition	**#lines**	**#wordsArabic/English**
Supplied BTEC	19,847	157,795/189,861
Full BTEC	39,511	305,272/361405
Travel books	31,388	English: 255,534

Table 2: *Training corpora for Arabic → English translation systems*

For the language models the English side of the supplied data was used for the Open data track; for the C-STAR data

track the English part of the Full BTEC corpus plus additional data from travel phrase books.

4.3.2. Official Submissions

Open data track Submission		
TM data	Supplied data only	
TM type	LogLin	
LM data	Supplied data only	
Scores (BLEU & NIST)		
ASR output	**0.1995**	**5.3359**
CRR	**0.2208**	**5.9059**
C-STAR data track Submission		
TM data	Full BTEC	
TM type	LogLin	
LM data	Full BTEC + travel books	
Scores (BLEU & NIST)		
ASR output	**0.2123**	**5.8693**
CRR	**0.2420**	**6.4073**

Table 3: *Official Submissions Arabic → English*

For the submitted system to the Open data track only the Supplied data was used. For the C-STAR data track we added the additionally translated BTEC data.

We can generally see that the ASR system introduces errors that affect the performance of the translation system. It is nice to note that the additionally translated data is able to boost the translation performance and the system achieves significantly better scores.

4.3.3. Contrastive Results

The contrastive results for Arabic → English show that the LogLin phrase extraction method usually outperforms the PESA phrase extraction technique on this task. We again see considerable improvements going from the Supplied data to the Full BTEC data.

Supplied data				
	PESA		LogLin	
Dev Set(ASR)	0.2275	5.8225	0.2430	5.8634
Dev Set (CRR)	0.2455	6.2317	0.2720	6.5101
Test Set (ASR)	0.1908	5.3794	0.1995	5.3359
Test Set (CRR)	0.2080	5.8344	0.2208	5.9059
Full BTEC + any available data (C-STAR track)				
	PESA		LogLin	
Dev Set(ASR)	0.2380	6.0998	0.2657	5.8690
Dev Set (CRR)	0.2696	6.6108	0.2864	6.8919
Test Set (ASR)	0.1989	5.6162	0.2123	5.8693
Test Set (CRR)	0.2138	6.0427	0.2420	6.4073

Table 4: *Contrastive Results for Arabic → English*

4.4. Italian → English

4.4.1. Training Corpora

For Italian → English 20,000 lines of data were provided. In addition we were able to gather 2,777 lines of Italian/English travel phrases from the web and added this to the data for the Open data track. The Full BTEC corpus for Italian has 55,413 lines.

Data Condition	**#lines**	**#wordsItalian/English**
Supplied	19,972	140,695/153,066
Web data (travel phrases)	2,777	11,748/ 13,345
Full BTEC	55,413	395,467/432,085

Table 5: *Training corpora for Arabic → English translation systems*

4.4.2. Official Submissions

For Italian → English it is interesting to note that the difference between the ASR output translation and the translation of the correct recognition results is much larger than for the other language pairs, partly due to the low word accuracy for the Italian 1-best ASR output of only 70.88%. Here we see a very significant drop of about 0.07 BLEU in performance.

There is also a significant difference between the Open data track and the C-STAR data track scores.

Open data track Submission		
TM data	Supplied and web data	
TM type	PESA	
LM data	Supplied and web data	
Scores (BLEU & NIST)		
ASR	**0.2388**	**6.1999**
CRR	**0.3030**	**7.3011**
C-STAR data track Submission		
TM data	Supplied + Full BTEC + web data	
TM type	PESA	
LM data	Supplied + Full BTEC + web data	
Scores (BLEU & NIST)		
ASR	**0.2630**	**6.6617**
CRR	**0.3312**	**7.7622**

Table 6: *Official submissions Italian → English*

4.4.3. Contrastive Results

The contrastive results show that the LogLin phrase extraction method is able to significantly outperform the PESA method. This is particularly true for the test set translations where we see an improvement of about 0.03-0.04 BLEU and far less pronounced for the translations of the development set.

Supplied data + free data (Open data track)				
	PESA		LogLin	
Dev Set (CRR)	0.3753	8.1078	0.3794	8.2301
Test Set (ASR)	0.2388	6.1999	0.2719	6.6064
Test Set (CRR)	0.3030	7.3011	0.3353	7.6730
Full BTEC + any available data (C-STAR track)				
	PESA		LogLin	
Dev Set (CRR)	0.4096	8.5651	0.4122	8.5923
Test Set (ASR)	0.2630	6.6617	0.2912	7.0812
Test Set (CRR)	0.3312	7.7622	0.3626	8.1408

Table 7: *Contrastive Results for Italian → English*

4.5. Chinese → English

4.5.1. Training Corpora

For Chinese → English 40,000 lines of data were supplied. The Full BTEC data for this language pair is complete at about 160,000 sentences. In addition we added 106,826 lines of data that was gathered from freely available bilingual newswire data using the test set of IWSLT 2005 as a query. We applied the technique described in [12]. Also the monolingual travel books were added to the language model for the C-STAR data track.

Data Condition	**#lines**	**#wordsChinese/English**
Supplied	39,953	351,060/306,149
Full BTEC	163,326	1,008,568/954,591
IR data	106,826	1,838,597/1,871,748
Travel books	31,388	English: 255,534

Table 8: *Training corpora for Chinese → English translation systems*

4.5.2. Word segmentation

Different word-segmentations had to be applied for the Open and C-STAR data track systems. For the Open data track system, the word-segmentation provided in the supplied data was used. However we could not duplicate this word segmentation for the C-STAR data track so all training, test and development corpora had to be re-segmented using our own segmenter. The translation scores indicate that the inferior word segmentation had an impact on the performance of the Chinese C-STAR data track system. This was also observed for the Japanese → English system.

4.5.3. Official submissions

This is the reason why the official submissions for the Chinese → English translation task showed the strange behavior that the score did not really improve when going from the Open data track to the C-STAR data track.

Open data track Submission		
Bilingual training data	Supplied Data only	
TM type	LogLin	
LM data	Supplied Data only	
Scores (BLEU & NIST)		
ASR spont.	**0.1630**	**4.9732**
ASR read	**0.1710**	**5.0768**
CRR	**0.1996**	**5.7603**
C-STAR data track Submission		
TM data	Full BTEC + IR data	
TM type	PESA	
LM data	Full BTEC + Travel Books + 4xSupplied_Data	
Scores (BLEU & NIST)		
ASR spont.	**0.1622**	**5.1865**
ASR read	**0.1645**	**5.2372**
CRR	**0.2057**	**6.0548**

Table 9: Official submissions Chinese → English

4.5.4. Contrastive Results

For the contrastive results we first investigated the impact of the LogLin phrase alignment model. The results in table 10 clearly show that for the Supplied Data and BTEC data the LogLin model shows consistent improvements, especially regarding BLEU score.

Supplied data				
	PESA		LogLin	
Test Set (ASR spont.)	0.1393	4.8752	0.1630	4.9732
Test Set (ASR read)	0.1539	5.0913	0.1710	5.0768
Test Set (CRR)	0.1846	5.8397	0.1996	5.7603
Full BTEC data				
	PESA		LogLin	
Test Set (ASR spont.)	0.1388	4.8686	0.1531	4.9926
Test Set (ASR read)	0.1436	5.0036	0.1894	5.7431
Test Set (CRR)	0.1737	5.7002	-	-

Table 10: *Contrastive Results for Chinese → English*

For the data conditions with additional data we did not apply the LogLin model but only used the PESA alignment. Table 11 shows results with added free and proprietary data. In both cases the BLEU scores increase significantly. It is not unexpected that additional travel books as in-domain data help in this situation. But even the out-of-domain IR data helped considerably.

4.6. Japanese → English

4.6.1. Training Corpora

Initially, three sets of training data were prepared for the Japanese → English systems as described in Table 12. The supplied data condition consists of 39,953 sentence pairs. For the C-STAR track submission system, two separate cor-

Supplied data + free data (Open data track)		
	PESA	
Test Set (ASR spont.)	0.1501	4.8736
Test Set (ASR read)	0.1654	5.0799
Test Set (CRR)	0.1972	5.8169
Full BTEC + any available data (C-STAR track)		
	PESA	
Test Set (ASR spont.)	0.1622	5.1865
Test Set (ASR read)	0.1645	5.2372
Test Set (CRR)	0.2057	6.0548

Table 11: *Contrastive Results for Chinese → English using additional data*

pora were combined; the Full BTEC corpus (consisting of 162,318 sentence pairs), and a small *medical-dialogs* corpus (4,019 sentence pairs). The *medical-dialogs* corpus was collected at InterACT and consists of bilingual Japanese/English spoken dialogues in the medical domain. This data was appended twice to the Full BTEC corpora to obtain a total training set of 170,356 sentence pairs.

Data Condition	**#lines**	**#words Japanese/English**
Supplied	39,953	403,323 / 381,776
Full BTEC	162,318	1,185,129 / 1,226,490
Medical dialogs	4,019	65,604 / 53,022

Table 12: *Training Corpora for Japanese → English Translation Systems*

Table 13 shows the translation performance of the provided development set for a system trained using the Full BTEC corpora and when the additional *medical-domain* corpus was incorporated. On the development set translation performance (both BLEU and NIST) were improved for both the ASR-output and the correct recognition result cases by incorporating the *medical-domain* corpora. This system was used as the C-STAR data track submission system.

Training Corpora	**Translation Performance**			
	ASR Output		CRR	
Full BTEC	0.1938	5.5076	0.2222	6.2152
Full BTEC + *medical-dialogs*	0.2001	5.5914	0.2289	6.2775

Table 13: *Translation performance for C-STAR-track system with additional data*

4.6.2. Word segmentation

Analogous to the Chinese → English system different word segmentations were used for the Open and C-STAR data track systems. For the Open data track system, the word segmentation provided in the supplied data was used. For the C-STAR data track, however, word segmentation was performed on all corpora using MeCab [13]. It can again be observed that this word segmentation is worse than the provided segmentation as the scores for the C-STAR data track do not significantly improve. Experiments showed that the performance of the MeCab segmentation was consistently about 0.01-0.02 BLEU lower than the performance of a system trained using the provided segmentation.

Furthermore, it was observed that re-segmenting the ASR output using a context-based segmenter (in this case MeCab) generated additional segmentation errors. On the development set, in addition to segmentation errors due to ASR errors, an additional 114 characters were incorrectly segmented due to the propagation of errors during context-based parsing. In future work, we intend to overcome such errors by considering all possible word segmentations during phrase matching.

4.6.3. Official Submissions

Four separate Japanese → English translation systems were developed for the IWSLT 2006 evaluation. For the Open track, systems were developed using the supplied data only (40k sentence pairs) and for the C-STAR track the Full BTEC and *medial-dialogs* corpora were used.

The translation performance of the systems submitted to the Open and C-STAR data tracks are shown in Table 14. Our Open data track system obtained a NIST score of 5.63 on ASR output, the second highest NIST score for this evaluation-track. Although adding the *medical-dialogs* corpora to our submission system improved performance on the development set, it caused a small degradation on the submission test set, suggesting that a larger and more homogenous development set may be required.

Open data track submission		
TM data	Supplied data only	
TM Type	PESA	
LM Data	Supplied data only	
Scores (BLEU & NIST)		
ASR	**0.1868**	**5.6343**
CRR	**0.2030**	**5.9322**
C-STAR data track submission		
TM data	Full BTEC + 2xmedical dialogs	
TM type	PESA	
LM data	Full BTEC + 2xmedical dialogs	
Scores (BLEU & NIST)		
ASR	**0.1841**	**5.3980**
CRR	**0.2007**	**5.8584**

Table 14: *Official submissions Japanese → English*

4.6.4. Contrastive Results

The contrastive results support this hypothesis (Table 15 and Table 16). We also see slight drops in the scores going from the supplied data to the Full BTEC data conditions. (see following section for a closer analysis)

The LogLin model shows little improvement for the development set on the Supplied Data but drops insignificantly on the test sets.

The difference between the translation of the ASR output and the correct recognition result is rather small here with a an improvement of about 0.02 (BLEU).

Supplied data				
	PESA		LogLin	
Dev Set (ASR)	0.2026	5.7974	0.2131	5.7787
Dev Set (CRR)	0.2325	6.4324	0.2434	6.4009
Test Set (ASR)	0.1868	5.6343	0.1830	5.5749
Test Set (CRR)	0.2030	5.9322	0.2009	5.6201

Table 15: *Contrastive Results for Japanese → English*

Full BTEC data		
	PESA	
Dev Set (ASR)	0.1938	5.5076
Dev Set (CRR)	0.2222	6.2152
Test Set (ASR)	0.1850	5.4349
Test Set (CRR)	0.2045	5.8719
Full BTEC + any available data (C-STAR track)		
	PESA	
Dev Set (ASR)	0.2001	5.5914
Dev Set (CRR)	0.2289	5.3980
Test Set (ASR)	0.1841	5.3980
Test Set (CRR)	0.2007	5.8584

Table 16: *Contrastive Results for Japanese → English*

5. Analysis of the results

The most surprising result about the previous scores is that for Chinese → English and Japanese → English the Full BTEC corpus does not really seem to help the translation performance. As mentioned before this is mainly due to differences in word segmentation. The provided data was segmented with manual support thus achieving a great segmentation quality. The Full BTEC corpora for Chinese and Japanese were not available in this same segmentation so a re-segmentation was necessary using standard segmentation tools.

Another reason is that looking at the data sizes we can note that only 20,000 lines of data were provided for Italian → English and Arabic → English while 40,000 lines of data were provided for Chinese → English and Japanese → English.

For Japanese → English and Chinese → English the Full BTEC corpus of about 160,000 lines is available while for Italian → English only about 55,000 lines were translated and overall only about 40,000 lines of BTEC are actually available for Arabic → English.

However the 20,000 lines of English data that were additionally translated to Arabic were carefully chosen to be very "informative" and not to contain too much repetition given the already available translations. This was done using the technique presented in [14]. The 55,000 lines of Italian → English data were produced using a similar method.

This means that the bare number of lines underestimates the impact of these additional translations in comparison to the impact the Full BTEC corpus of 160,000 lines would have.

This means that the relative increase in training data (especially useful training data) is probably larger for the Italian → English and Arabic → English systems than for the Japanese → English and Chinese → English systems.

Concerning the two investigated phrase extraction methods PESA and LogLin we notice the general trend that LogLin gives a better performance than PESA. But the main problem with LogLin applied in this task is that it needs hand-aligned training data from the same language pair and to a lesser extent from the same domain as the testing data. These *gold-blocks* are necessary to tune the alignment process to the specifics of a given language pair. The only hand-aligned data that was available is newswire data for the language pair Chinese → English. If it would be possible to use actual BTEC data here we would expect to see greater improvements. We tried the LogLin aligment on other language pairs anyway and as the results show we still got significant improvements on Italian and Arabic while the drop on Japanese was not significant. It is especially surprising to see an improvement of over 0.03 BLEU on Italian-English. Figure 1 compares the scores for the source languages Arabic, Italian, Chinese and Japanese (Test set (ASR), Open data track).

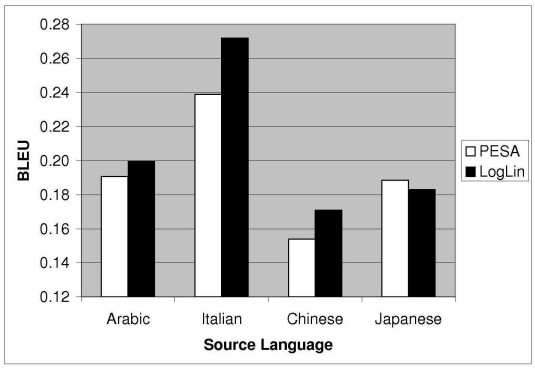


Figure 1: *LogLin improvements vs. PESA*

6. Future Work

The focus of the IWSLT 2006 translation campaigns has traditionally been speech-to-speech translation using ASR output or even speech as the input of a speech-to-speech translation system. In this years evaluation campaign we did not use any specific approaches to translate ASR output compared to translating correct recognition results as we just translated the 1-best translation.

It has however been shown that using an n-best list or a lattice as the input for a translation system can give better results in this situation as the lattice might still contain potentially better paths [15]. For the lattices in this evaluation this is especially true for Arabic where the word accuracy of the best path in the lattice is 88.20% compared to 73.88% for the 1-best word accuracy with similar situations for the other language pairs. One of our future goals will be to use this additional information in form of lattices or n-best lists during the translation process.

7. Acknowledgements

The work reported here was partly funded by the European Union (EU) under the integrated project TC-STAR (Grant number IST-506738), by the National Science Foundation (NSF) under the integrated project STR-DUST (Grant number IIS-0325905) and by stipends from the InterACT exchange program.

8. References

[1] Kolss, Muntsin, Zhao, Bing, Vogel, Stephan, Hildebrand, Almut Silja, Niehues, Jan, Venugopal, Ashish and Zhang, Ying, The ISL Statistical Machine Translation System for the TC-STAR Spring 2006 Evaluation, in TC-STAR Workshop on Speech-to-Speech Translation, TC-STAR-WS 2006, Barcelona, Spain, June 2006.

[2] Hewavitharana, Sanjika, Zhao, Bing, Hildebrand, Almut Silja, Eck, Matthias, Hori, Chiori, Vogel, Stephan and Waibel, Alex, The CMU Statistical Machine Translation System for IWSLT 2005, in Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005), Pittsburgh, USA, October, 2005

[3] Vogel, Stephan, PESA: phrase pair extraction as sentence splitting, in Proceedings of the Machine Translation Summit X, Phuket, Thailand, September 2005.

[4] Brown, Peter F., Della Pietra, Vincent J., Della Pietra, Stephen A. and Mercer, Robert L., The mathematics of statistical machine translation: Parameter estimation, Computational Linguistics, vol. 19, no.2, 1993.

[5] Zhao, Bing and Vogel, Stephan, A generalized alignment-free phrase extraction, in Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, Michigan, June 2005.

[6] Zhao, Bing and Waibel, Alex, Learning a log-linear model with bilingual phrase-pair features for statistical machine translation, in Proceedings of the SigHan Workshop, Jeju, Korea, October 2005.

[7] Zhang, Ying, Hildebrand, Almut Silja and Vogel, Stephan, Distributed Language Modeling for N-best List Re-ranking, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP 2006), Sydney, Australia, July 2006.

[8] Vogel, Stephan, SMT Decoder Dissected: Word Reordering, in Proceedings the International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, October 2003.

[9] Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, WeiJing. 2002. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL 2002), Philadelphia, USA, July 2002.

[10] Doddington, George, Automatic Evaluation of machine translation quality using n-gram co-occurrence statistics, in Proceedings of the Human Language Technology Conference (HLT 2002), San Diego, USA, March 2002.

[11] Takezawa, Toshiyuki, Sumita, Eiichiro, Sugaya, Fumiaki, Yamamoto, Hirofumi. Towards a broad-coverage bilingual corpus for speech translations of travel conversations in the real world. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Spain, May 2002.

[12] Hildebrand, Almut Silja, Eck, Matthias, Vogel, Stephan and Waibel, Alex, Adaptation of the translation model for statistical machine translation based on information retrieval, in Proceedings of the EAMT 2005 (European Association for Machine Translation, Budapest, Hungary, May, 2005.

[13] http://mecab.sourceforge.jp/, MeCab Japanese segmenter

[14] Eck, Matthias, Vogel, Stephan and Waibel, Alex, Low Cost Portability for Statistical Machine Translation based on N-gram Coverage, in Proceedings of the Machine Translation Summit X, Phuket, Thailand, September 2005.

[15] Saleem, Shirin, Jou, Szu-Chen, Vogel, Stephan and Schultz, Tanja, Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems, in Proceedings of the International Conference of Spoken Language Processing (ICSLP-2004), Jeju Island, South Korea, October 2004.