IBM Statistical Machine Translation for Spoken Languages

Young-Suk Lee

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

Abstract

We discuss performance enhancing techniques we have developed for the *IWSLT 2005 Evaluation Campaign*: (i) a phrase acquisition technique which expands the phrase boundaries to include target words aligned to null source words in a principled manner, and (ii) a system combination technique which selects the minimum cost translation output out of many translation outputs of the same input segment produced by various systems using different phrase translation lexicons. We also discuss IBM system performances in the Arabic to English and Chinese to English translation evaluations of the IWSLT 2005 evaluation campaign.

1. Introduction

IBM spoken language translation system is based on a statistical translation model introduced in [1]. We adopt a phrase translation model as the baseline, [2], [3], [4], [5], [6]. We improve the baseline system performance by an extended phrase selection algorithm and a novel system combination technique.

The baseline phrase selection algorithm, cf. Section 2.1, is augmented by a technique which expands a target phrase to include target words typically aligned to null source words in the neighborhood of a high precision word alignment (and possibly a target words with an explicit alignment link next to it). The expanded phrase expansion algorithm is effective for language pairs in which one source word should be aligned to many target words in principle.

For system combination, we produce translation outputs of the same input segment with various systems, which use different phrase translation lexicons. Then we select the output with the minimum translation cost. The idea behind is to capitalize on the strength of each phrase translation lexicon even though the system using one particular lexicon may generally result in the highest translation quality. For instance, a phrase translation lexicon derived from a small domain-specific training corpus might produce a higher translation quality on the evaluation corpus from the same domain than a lexicon derived from a large domain neutral training corpus. And yet, there can be some segments of the evaluation corpus which are better translated by the large domain neutral lexicon than by the small domain-specific lexicon. If some rare words in the small domain-specific corpus occur frequently in the large domain neutral corpus, they are likely to be better translated by the system using the large lexicon than by the system using the small domain-specific lexicon.

In Section 2, we give an overview of the baseline phrase translation system. In Section 3, we discuss the extended phrase extension algorithm and the system combination technique. In Section 4, we show the impact of various techniques on Arabic to English and Chinese to English translations. Finally in Section 5, we discuss our ongoing work.

Throughout this paper, we use the term *block* (*b*) to denote a phrase translation pair consisting of a source ($\bar{f}$) and a target phrase ($\bar{e}$).

2. Baseline Phrase Translation System

We discuss the acquisition of phrase translation lexicon and the phrase decoder below, which have been developed for Spanish-English translations under the project TC-STAR.

2.1. Acquisition of Phrase Translation Lexicon

Phrase translation lexicon is obtained via word alignment and block selection algorithms.

We obtain word alignment between source and target language sentences by application of HMM alignments [8]. We word-align a parallel corpus bi-directionally: 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: $e \rightarrow f$), where f denotes a source word position and e a target word position. We define precision (A_P) and recall (A_R) oriented alignments as follows:

$$A_P = A_1 \cap A_2$$
$$A_R = A_1 \cup A_2$$

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. The set of all source word positions covered by some word links in A are denoted as $col(A)$.

Starting from a high precision word alignment A_P, we obtain blocks according to the projection and extension algorithms [5]. **Projection Algorithm:** We first identify source intervals $[f, f']$, where $f, f' \in col(A_P)$. Then we compute the minimum target index e and the maximum target index e' for the word links that fall in the interval $[f, f']$:

$$[f, f'] \rightarrow [\min_{e \in Pf([f,f'])} e, \max_{e' \in Pf([f,f'])} e']$$

$P_f(\cdot)$ denotes the projection from source intervals into target intervals. The block consisting of the target and source words at the link positions is denoted as b. **Extension Algorithm:** We expand the alignment links to include alignment points in the neighborhood of the high precision alignment A_P and lie within the high recall alignment A_R. The extensions are carried out iteratively until no new alignment links from A_R are added. See [5] for the details of the extension algorithm.

Once the blocks are collected according to the projection and the extension algorithms, one-one blocks consisting of one source word and one target word are further derived from the word alignment A_2, which aligns from a target word position to a source word position. In addition, blocks containing non-consecutive source word sequence are filtered out.

2.2. Decoding

Our phrase decoder utilizes 10 distinct scoring functions (listed below) multiplied by their respective weight:

- Direct phrase translation model cost
- Source-channel phrase translation model cost
- Unigram phrase translation model cost
- IBM Model 1 cost applied in both directions
- Language model cost at phrase boundaries
- Language model cost within a phrase
- Word & block count penalties
- Outbound and inbound distortion model costs

Direct phrase translation model probabilities are obtained according to (1).

$$(1)\ p(\bar{e} \mid \bar{f}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{e}'} count(\bar{e}', \bar{f})}$$

Source-channel model probabilities are computed, according to (2).

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

Unigram translation probability of a block ($b = \bar{e}, \bar{f}$), is obtained according to (3):

$$(3)\ p(b) = \frac{count(b)}{\sum_{b'} count(b')}$$

IBM Model 1 translation cost is computed for each phrase, as in (4):

$$(4)\ \sum_{j=1}^{m} -\log 10 \max p(f_j|e_i), 1 \le i \le n$$

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. For the current evaluations, we set the value of β from 3.5 to 5.0. Model 1 translation probability $p(f_j \mid e_i)$ is computed by the relative frequency of one-one blocks (i.e. blocks consisting of one source word and one target word), as in (5).

$$(5)\ p(f \mid e) = \frac{count(f, e)}{\sum_{f'} count(f', e)}$$

Trigram language model probabilities are obtained for each word in the target phrase ($\bar{e}$), according to (6).

$$(6)\ p(e_i \mid e_{i-1}, e_{i-2})$$

Weight assigned to the word trigram language model probability may be set differently for the first word e_1 and the remaining word e_i ($1 < i < n$, n: number of words in the target phrase).

Word and block count penalties are applied to ensure that the decoder does not always choose the shortest translation output and the longest matching source phrase, respectively, analogous to those in [13]. Word level distortion models in [21] are incorporated into the decoder, as well.

3. Performance Enhancing Techniques

We now discuss two core performance enhancing techniques: extended block extension algorithm and system combination.

3.1. Extended Block Extension Algorithm

We have extended the block extension algorithm in 2.1 to capture the asymmetrical properties of word and

sentence structures often present between the source and target languages.

Given the Arabic sentence in Buckwalter transliteration (7a) and its English translation (7b), we typically obtain the word alignment in (8), where the target words *i*, *do*, *it* are not aligned to any source words:

(7) a. lA Aryd AzAlthA .
b. i do n't want it extracted .

(8) lA <==> n't
Aryd <==> want
AzAlthA <==> extracted
. <==> .

Similarly, given the Chinese sentence (9a) and its English translation (9b), we typically obtain a word alignment, as in (10), where the target words *is* and *the* are not aligned to any source words:

(9) a. 早餐 多少 钱 ?
b. how much is the breakfast ?

(10) 早餐 <==> breakfast
多少 <==> how
钱 <==> much
? <==> ?

Since the missing word alignment, as in (8) and (10), is due to intrinsic asymmetries between the source and the target language grammars, rather than deficiencies of word alignment itself, we incorporate these asymmetries into the block extension algorithm as follows: ***First***, we collect the list of target words typically aligned to a null source word, e.g. *i*, *it*, *do* in Arabic-English, and *is*, *the* in Chinese-English, and call them **expansion word list**. ***Second***, in applying the block extension algorithm, if a target word is included in the expansion word list, and occurs in the neighborhood of high precision word alignment A_P, we extend the target phrase to include the expansion word even if there is no alignment link between the target and any source word positions in either alignment direction. (11) shows the expansion word list we used for Arabic to English translation without word segmentation into morphemes, and (12) is the expansion word list for Arabic to English translation with morphological analysis [7].

(11) *i, you, is, to, are, am, do, does, it, 's, 'll, 'm, 're, a, an, the, and, will, 'd, be, me, him, her, them, us, my, your, his, her, their, its, our, mine, yours, hers, theirs, ours, two, for, with, been*

(12) *i, you, is, to, are, am, do, does, it, 's, 've, 'll, n't, 'm, 're, 'd, be, your, my, two, any, some, we, they, the, been, and, for, with, that, would, his, her, most*

We also extend the block extension algorithm to include a target word if the target word position is aligned to a source word position by an intersection relation, and is adjacent to another target word position with no alignment link. The extended block extension algorithm enables us to obtain two more blocks (b) and (c) extended from the seed block (a) in Figure 1.

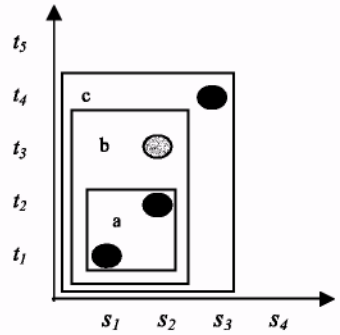

Figure 1. Blocks acquired by the Extended Block Extension Algorithm

In Figure 1, solid circles denote the word link by an intersection relation, and the grey circle denotes a target word without an alignment link in the neighborhood of the seed block at a possible extension point. $s_1 ... s_4$ denote source word positions, and $t_1 ... t_5$ target word positions. The seed block (a) is obtained from the high precision alignment A_P. The extended block extension algorithm allows the target phrase to be extended to include the null link target word t_3, leading to the extended block (b). By a recursive application of the block extension algorithm, we further obtain the extended block (c), containing the sequence of target words without and with an alignment link.

3.2. System Combination

A procedure for system combination is given below:

Step 1: Build various types of phrase translation lexicons, which may vary according to their vocabulary coverage and/or phrase selection algorithms.

Step 2: Translate the same segment (roughly equivalent to a sentence) and compute its translation cost with systems using each of the lexicons acquired in Step 1.

Step 3: Compare the translation costs attributed to systems using each of the phrase translation lexicons and choose the translation output of the system which results in the minimum cost according to the algorithms in Section 3.3.3.

An example of two phrase translation lexicons differing from each other in terms of vocabulary coverage in **Step 1** would be (i) one built from a small domain-specific training corpus, and (ii) the other built from a large domain-neutral training corpus. Another example from Chinese to English translation can be (i) one built from character-segmented Chinese corpus, and (ii) the other built from a word-segmented Chinese corpus. Two phrase translation lexicons can also differ from each other in terms of distinct algorithms by which each of the lexicons are derived even if the training corpus is the same.

Below we discuss system combinations for Arabic to English and Chinese to English translations.

3.3.1. Arabic to English Translation

An Arabic word typically corresponds to more than one English word. For instance, an Arabic word *AlAHmr* corresponds to two English words *the red*, and *llmEArDp* to three English words *of the opposition*. This rich morphology of Arabic induces a data sparseness problem. A way of overcoming the data sparseness problem using the same training corpus is to segment an Arabic word into a morpheme sequence of *prefix-stem-suffix*, as described in [11], e.g. *llmEArDp* → *l l mEArD p*, so that an Arabic morpheme roughly corresponds to an English word. We can accomplish an even stronger morphological symmetry between an Arabic morpheme and an English word by merging an Arabic prefix/suffix into its stem or deleting a prefix/suffix, e.g. *l l mEArD p* → *l l mEArDp*, so-called morphological analysis [7].

On the basis of the observation above, we derive three kinds of phrase translation lexicons from three different types of word alignments, differentiated by Arabic corpus processing, as in (13):

(13) a. Un-segmented Arabic
b. Word-segmented Arabic (prefix-stem-suffix)
c. Morphologically analyzed Arabic

The English corpus is uniformly punctuation tokenized for all three types of word alignments. OOV ratio of various evaluation corpora on the three phrase translation lexicons (derived from the 20K sentence pairs for the supplied data track) is shown in Table 1.

Arabic corpus	IWSLT05	IWSLT04	CSTAR03
un-segmented	5.0 %	4.68 %	4.58 %
	159/3164	151/3226	144/3146
word-segmented	1.4 %	1.1 %	1.35 %
	59/4747	53/4774	62/4594
morphological analysis	2.1 %	2.1 %	2.14 %
	83/3914	83/3940	82/3837

Table 1. OOV ratio of various phrase lexicons

For the unrestricted data condition, we additionally apply the system combination to a system using a phrase translation lexicon derived from a large training corpus containing both the domain specific and out-of-domain corpora. Table 2 shows the OOV ratio of the phrase translation lexicons derived from 20K sentence pair supplied training corpus and 143,253 sentence pair unrestricted training corpus with un-segmented Arabic, which contains both the supplied data and LDC-distributed Arabic-English news and multiple translation corpora.

corpus size	IWSLT05	IWSLT04	CSTAR03
20K supplied	5.0 %	4.68 %	4.58 %
	159/3164	151/3226	144/3144
143K unrestricted	2.21 %	2.08 %	2.04 %
	70/3164	67/3226	64/3144

Table 2. OOV ratio of lexicons derived from domain-specific and domain-neutral training corpora

3.3.2. Chinese to English Translation

Analogous to Arabic to English translation, we perform system combination across phrase translation lexicons of varying vocabulary coverage, derived from word alignments differentiated by Chinese corpus processing, as in (14):

(14) a. Word-segmented Chinese
b. Character-segmented Chinese

For Chinese word segmentation (14a), we use a language model-based Chinese word segmentation system [10], trained on the 20K supplied data corpus distributed for IWSLT 2004 evaluations. For character segmentation of Chinese, we segment the corpus at each character position. OOV ratio of evaluation corpora on the lexicons derived from word- and character-segmented Chinese are shown in Table 3.

Chinese corpus	IWSLT05	IWSLT04	CSTAR03
word-segmented	2.15 %	2.12 %	2.9 %
	82/3822	76/3578	102/3510
character-segmented	0.51 %	0.19 %	0.55%
	26/5128	9/4749	26/4715

Table 3. OOV ratio of lexicons derived from word- and character-segmented Chinese corpora

We also apply reordering rules, (15), (24), to word- and character-segmented Chinese, resulting in lexicons of different translation qualities but with the same vocabulary coverage. We apply the system combination to systems using lexicons derived from re-ordered and un-reordered Chinese corpora [23].

(15) **Chinese re-ordering rules**

a. Word re-ordering: Identify Chinese question words corresponding to English *where*, *which*, *who*, *when*, *how*, *what*, *why* in a Chinese question sentence. Move the Chinese word sequence starting from the question word to the one before the sentence ending markers (了, 吗, ？) to the beginning of the sentence.

b. Character re-ordering: Move the last three characters of a Chinese question sentence excluding the sentence ending markers (了, 吗, ？) to the beginning of the sentence in the same order.

3.3.3. System Combination Algorithm

We first determine the system using the phrase translation lexicon which generally results in the highest translation quality measured by BLEU [9]. We call the system producing the highest translation quality *h-sys*, and the systems producing the lower translation qualities, $l\text{-}sys_1$, $l\text{-}sys_2$, …, $l\text{-}sys_n$. Other notations are given below:

- cost(system): translation cost of the system output computed by the system's decoder
- tarlen(system): target length produced by the system's decoder
- srclen(system): input segment length to be translated by the system
- oov(system): number of out-of-vocabulary of the input segment computed at each source word position on the basis of the system's lexicon
- output(system): system's output

System combination algorithms are given below:

Arabic to English

For each input segment of the evaluation corpus:

if cost(*h-sys*) > cost($l\text{-}sys_1$) + *threshold_1* &
  oov(*h-sys*) = 0
  then choose output($l\text{-}sys_1$)
…
else if
  cost(*h-sys*) > cost($l\text{-}sys_n$) + *threshold_n* &
  oov(*h-sys*) = 0,
  then choose output($l\text{-}sys_n$)
else
  choose output(*h-sys*)

The algorithm states that if the translation cost of the highest-performing system *h-sys* is higher than that of a lower-performing system $l\text{-}sys_n$ by a specified threshold *threshold_n*, then choose the translation output of $l\text{-}sys_n$ over that of *h-sys*. When there are more than one lower performing systems, i.e. $n > 1$ in $l\text{-}sys_n$, the order in which the comparison is made with output(*h-sys*) is determined on the basis of the effectiveness of the system combination between *h-sys* and $l\text{-}sys_n$, i.e. which two system combinations give rise to a higher BLEU score. The more effective the system combination is, the earlier the comparison is made. Note that the output selection is greedy since once the selection is made, it is definite and the comparison for the given input segment stops. The values *threshold_1*, …, *threshold_n*, are manually set to those which result in the highest performance improvement, and typically vary according to the properties of two systems for which the comparisons are made.

Chinese to English

For each input segment in the evaluation corpus:

if cost(*h-sys*) > cost($l\text{-}sys_1$) + *threshold_1* &
  oov(*h-sys*) = 0 & tarlen($l\text{-}sys_1$) > tarlen(*h-sys*) / 2
  then choose output($l\text{-}sys_1$);
…
else if
  cost(*h-sys*) > cost($l\text{-}sys_n$) + *threshold_n* &
  oov(*h-sys*) = 0 & tarlen($l\text{-}sys_n$) > tarlen(*h-sys*) / 2
  then choose output($l\text{-}sys_n$);
else
  choose output(*h-sys*);

A major difference between Arabic to English and Chinese to English algorithms is the target output length imposed on Chinese to English translation. For Chinese to English, the tarlen($l\text{-}sys_n$) should be longer than half of the tarlen(*h-sys*) for output($l\text{-}sys_n$) to be selected. This is to counteract the decoder tendency of producing lower translation cost for excessively short translation output.

We now point out a couple of issues that can be further improved.

First, effectiveness of system combination largely depends on the value of *threshold_n*, which in turn depends on decoder parameter values. Currently, we learn this value by trying out several values whenever we change the decoder parameter setting. We are working to automatically identify the threshold values.

Second, currently the order in which the cost comparison is made between two systems is pre-determined and the output selection is greedy. We are trying to come up with a technique to select the minimum cost translation output on the basis of simultaneous comparisons of all systems to avoid incorrect greedy decisions.

Languages	Data tracks	BLEUr16n4	WER	PER	NIST	METEOR	GTM
A2E	**unrestricted**	**0.5996**	**0.3331**	**0.2939**	**9.7570**	**0.7261**	**0.6824**
	supplied+tools	0.5604	0.3565	0.3086	9.5922	0.7117	0.6664
	supplied	0.5384	0.3779	0.3363	8.6163	0.6887	0.6475
C2E	**unrestricted**	**0.4985**	**0.4337**	**0.3716**	**8.1719**	**0.6626**	**0.6103**
	supplied+tools	0.4785	0.4450	0.3792	7.8783	0.6513	0.5966
	supplied	0.4402	0.4692	0.3909	8.4357	0.6424	0.5878

Table 4. IBM Statistical Machine Translation System Performances in IWSLT 2005

4. Performance Evaluations

IBM system performances for Arabic to English (A2E) and Chinese to English (C2E) translations in the IWSLT 2005 evaluation campaign are shown in Table 4, where the data tracks we focused on are written in bold. Below we discuss the training corpora and the impact of new techniques discussed in Section 3 in terms of translation quality measured by BLEU.

4.1. Training corpora & tools

Training corpora according to evaluation conditions are given in Tables 5 & 6, where TM stands for parallel translation model training corpora by sentence pair count, and LM for the English language model training corpora by word count.

	Supplied	Unrestricted
TM	20k supplied	20k supplied 500 IWSLT 04 development set ~123k news (LDC)
LM	~190k supplied	~190k supplied ~191k JE supplied/IWSLT 04 ~1 b English Gigaword (LDC)

Table 5. Arabic to English training corpora

For supplied data condition, we trained Arabic to English system on un-segmented Arabic, and Chinese to English system on character-segmented Chinese. For supplied+tools and unrestricted data conditions, we have used Arabic word segmenter and morphological analyzer for Arabic to English, and Chinese word segmenter for Chinese to English translations.

	supplied	unrestricted
TM	20k supplied	20k supplied
LM	~190k supplied	~190k supplied ~191k JE supplied/IWSLT 04

Table 6. Chinese to English training corpora

4.2. Impact of extended block extension algorithm

Impact of extended block extension algorithm, cf. Section 3.1, on Arabic to English translation is shown in Tables 7 & 8, where the systems for supplied+tools condition are trained on morphologically analyzed Arabic corpus.

	Old algorithm	New algorithm
supplied	0.5503	0.5657
supplied+tools	0.5448	0.6102

Table 7. Impact of extended block extension algorithm on the A2E CSTAR 03 development test set

	Old algorithm	New algorithm
supplied	0.5312	0.5460
supplied+tools	0.5282	0.5723

Table 8. Impact of extended block extension algorithm on the A2E IWSLT 04 development test set

Impact of the extended block extension algorithm on Chinese to English translation is shown in Tables 9 & 10, where the systems for supplied+tools condition is trained on word-segmented Chinese corpus.

	Old algorithm	New algorithm
supplied	0.3343	0.3823
supplied+tools	0.3606	0.4243

Table 9. Impact of extended block extension algorithm on the C2E CSTAR 03 development test set

	Old algorithm	New algorithm
supplied	0.3608	0.4159
supplied+tools	0.3930	0.4418

Table 10. Impact of extended block extension algorithm on the C2E IWSLT 04 development test set

4.3. Impact of system combination

For Arabic to English translation, the system using the phrase translation lexicon derived from a morphologically analyzed Arabic corpus results in the highest BLEU score. Therefore, we take the system using the lexicon derived from morphologically analyzed Arabic as *h-sys*, cf. Section 3.3.3. Impact of system combination on IWSLT 2005 Arabic to English translation evaluations is shown in Table 11.

For the supplied+tools condition, two system outputs are combined: one using the lexicon derived

from morphologically analyzed Arabic and the other using the lexicon derived from word-segmented Arabic.

	supplied+tools	unrestricted
Morph analysis	0.5541	0.5862
Combination	0.5604	0.5996

Table 11. Impact of system combination on IWSLT 05 A2E evaluations

For the unrestricted data condition, three system outputs are combined: (i) system using the lexicon derived from morphologically analyzed Arabic and the supplied data, (ii) system using the lexicon derived from word-segmented Arabic and the supplied data, (iii) system using the lexicon derived from un-segmented Arabic and the unrestricted data.

Contribution of each system in the unrestricted data condition is shown in Table 12.

Systems	Segments Selected	
	Count	Ratio (%)
Morphological analysis	292	57.7
Word segmented	207	40.9
Unsegmented	7	1.38

Table 12. System selection statistics in IWSLT 05 A2E unrestricted data track

Impact of system combination on Chinese to English translation is shown in Table 13.

	supplied+tools	unrestricted
Word segmentation	0.4682	0.4874
Combination	0.4785	0.4985

Table 13. Impact of system combination on IWSLT 05 C2E evaluations

For Chinese to English translation, a phrase translation lexicon derived from word-segmented and reordered Chinese results in the highest BLEU score. Therefore, we take the system using the lexicon derived from word segmented and reordered Chinese as *h-sys*, cf. Section 3.3.3. For both supplied+tools and unrestricted data conditions, three systems are combined: (i) system using the lexicon derived from word-segmented and reordered Chinese, (ii) system using the lexicon derived from un-reordered character-segmented Chinese, and (iii) system using the lexicon derived from reordered character-segmented Chinese.

Contribution of each system in the unrestricted data track is shown in Table 14.

Systems	Segments Selected	
	Count	Ratio (%)
Word-segmented reordered	379	74.9
Char-segmented un-reordered	116	22.92
Char-segmented reordered	11	2.17

Table 14. System selection statistics in IWSLT 05 C2E unrestricted data track

5. Summary and Ongoing Work

We have discussed two key performance enhancing techniques we have newly developed for IWSLT 2005. First, extended phrase selection algorithm which enables the system to produce target phrases with null word alignment and its extension improves the performance statistically significantly on the CSTAR 03 and IWSLT 04 development test data for both Arabic to English and Chinese to English translations. Second, the system combination algorithm which combines the best translation output from various phrase translation systems is effective for both Arabic to English and Chinese to English translations. This technique is also effective for improving domain-specific translation quality by adding out-of-domain training corpora, which has been a challenge, cf. IWSLT 2004 evaluation results for Chinese to English unrestricted data track and supplied data track.

We are currently examining factors which cause the system combination algorithm to perform sub-optimally. This includes (i) automatic identification of threshold value for translation cost comparison, (ii) identification of parameters and features which have positive and negative impacts on system performances. We are also extending the system combination technique to other language pairs, e.g. English-Spanish, with a relatively large training corpora. We are also looking into word alignment techniques in order to improve the overall qualities of phrase translation lexicons.

Acknowledgements

This work has been funded in part by the European Commission under the project TC-STAR (Technology and Corpora for Speech to Speech Translation). We would like to thank Salim Roukos for comments on the earlier versions of this paper. We also would like to thank IBM Statistical Machine Translation Project Team on whose earlier work the current spoken language translation system is built.

References

[1] 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.

[2] 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.

[3] K. Yamada and K. Knight. "A syntax-based statistical translation model", *Proceedings of the 39th ACL–2001 Conference*, pages 523–530, 2001.

[4] D. Marcu and W. Wong. "A phrase-based, joint probability model for statistical machine translation", *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002.

[5] 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.

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

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

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

[9] 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.

[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] 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.

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

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

[14] D. Melamed, R. Green, and J. Turian. "Precision and recall of machine translation", *Proceedings of HLT–NAACL 2004*, 2004.

[15] G. Doddington. "Analysis of NIST Evaluation Data", NIST presentation at DARPA IAO Machine Translation Workshop, Santa Monica, CA, USA, July 22–23, 2002.

[16] M. Paul, H. Nakaiwa, and M. Federico. "Towards innovative evaluation methodologies for speech translation", *Working Notes of NTCIR–4*, Tokyo, 2–4 June 2004.

[17] Y-S. Lee, D. Sinder, and C. Weinstein. "Interlingua-based English-Korean two-way speech translation of doctor-patient dialogues with CCLINC", *Machine Translation*, 17(3):213–243, 2002.

[18] L. Levin, A. Lavie, M. Woszczyna, D. Gates, M. Gavaldà, D. Koll, and A. Waibel. "The Janus–III translation system: speech-to-speech translation in multiple domains", *Machine Translation*, 15:3–25. 2000.

[19] Y. Qu, B. DiEugenio, A. Lavie, L. Levin and C. P. Rose. "Minimizing cumulative error in discourse context", In *Dialogue Processing in Spoken Language Systems*, Springer Verlag, 1997.

[20] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech", *Computational Linguistics*, 26(3):339–374. 2000.

[21] Yaser Al-Onaizan, "IBM Site Report: Distortion-Based Word Reordering", *Proceedings of DARPA Machichine Translation Evaluation Workshop*, Alexandria, VA, USA, June 22–23, 2004.

[22] 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, ATR/Kyoto, Japan, September 30–October 1, 2004.

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

[24] Young-Suk Lee, "IBM Site Report: N-Best Reordering of Arabic for Statistical Machine Translation", *Proceedings of DARPA Machine Translation Workshop*, Alexandria, VA, USA, June 22–23, 2004.