IBM Spoken Language Translation System Evaluation

Young-Suk Lee, Salim Roukos

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

Abstract

We discuss phrase-based statistical machine translation performance enhancing techniques which have proven effective for Japanese-to-English and Chinese-to-English translation of BTEC corpus. We also address some issues that arise in conversational speech translation quality evaluations.

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, for which the unit of translation is a phrase consisting of one or more words, [2], [3], [4], [5], [6].

The baseline system is augmented by the morphological analysis detailed in [7] for an improved word alignment and phrase selection. System performance is significantly improved by phrase selection from recall oriented word alignments (see Section 2 for the definition) and filtering. Re-ordering of source language sentence into the target language word order, [21], [22], further improves phrase selection and word order accuracy. Non-monotone decoding and language model probability computation for every word in a target phrase enhances the translation quality over monotone decoding and language model probability computation only for words at phrase boundaries.

In Section 2, we give an overview of the baseline system. In Section 3, we discuss translation quality enhancing techniques along with experimental results. In Section 4, we address some issues in conversational speech translation evaluation. We discuss future work in Section 5.

We use the term *block* (*b*) to denote a phrase translation pair consisting of a source phrase ($\bar{f}$) and a target phrase ($\bar{e}$). We use the symbol $Pr(\cdot)$ to denote general probability distribution and $p(\cdot)$ to denote model-based probability distribution.

2. Baseline System Overview

Our baseline phrase translation system described in [Tillmann 2003] consists of three major components: word alignment, block selection, and decoding.

2.1. Word Alignment

We obtain word alignment between the source and the target language sentences by successive application of IBM Model 1 viterbi alignment for initialization and iterative HMM-based alignment, [8], for refinement.

We align a parallel corpus bi-directionally: one from the source language to the target language (A_1: $f \rightarrow e$) and the other from the target language to the source language (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_P are denoted as $col(A_P)$.

2.2. Block Selection

Starting from a high precision word alignment A_P, we obtain blocks according to (i) a projection algorithm and (ii) a block extension algorithm.

Projection Algorithm: We first project source intervals $[f', f]$, where $f', f \in col(A_P)$. We compute the minimum target index e' and maximum target index e for the word links that fall into the interval $[f', f]$:

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

$P_f(\cdot)$ projects source intervals into target intervals. The pair ($[f', f]$, $[e', e]$) defines a block alignment link *a*. The block consisting of the target and source words at the link positions is denoted as *b*. Target and source words in a block are subject to the contiguity condition.

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.

Among the candidate blocks obtained according to

the projection and extension algorithm, blocks satisfying the following three conditions are kept for use in translation:

i. Source phrase ($\bar{f}$) length ≤ 10 morphemes[1]
ii. Target phrase ($\bar{e}$) length ≤ 10 morphemes
iii. Block (b) frequency > 1

2.3. Decoding

Two types of model parameters, block unigram model and word trigram language model, are used in the baseline decoder. Block unigram probability is defined in (1), where n is the number of distinct blocks:

$$(1)\quad p(b) = \frac{count(b)}{\sum_{i=0}^{n} count(b_i)}$$

Word trigram probability is computed at target phrase boundaries only, skipping over words within a target phrase in case the target phrase length ≥ 2. Trigram language model probability between adjacent target phrases is computed, as in (2).

$$(2)\quad p(\bar{e}_i \mid \bar{e}_{i-1}) = p(e_1 \mid e_h, e_{h-1})$$

$\bar{e}_i$ is the current target phrase, $\bar{e}_{i-1}$ is the previous (one or more) target phrase in the hypothesis. e_1 is the first word of $\bar{e}_i$ [2], e_h the last target word in the hypothesis and e_{h-1} the second to the last target word in the hypothesis. The task of the decoder is to find the block sequence that maximizes the product of the unigram block probability and the trigram language model probability without reordering.

In decoder implementation, we use a DP-based beam search procedure. We start with an initial empty hypothesis. We maximize over all block segmentations $b_{1,n}$, where n is the number of blocks covering the input sentence, with the source phrases yielding a segmentation of the input sentence, generating the target sentence simultaneously. The decoder processes the input sentence 'cardinality synchronously', i.e. all partial hypotheses active at a given point cover the same number of input words. We prune out weaker hypotheses based on the cost (for block unigram probability and trigram language model probability) they incurred so far. The cheapest final hypothesis − the hypothesis with the highest probability − with no un-translated source words is the translation output.

[1] Morpheme is defined to be the minimal unit of meaning, and may or may not overlap with words, e.g. the Japanese object case marker を and English plural marker –*s* are a morpheme but not a word, whereas *president* in English これ 'this' in Japanese are both a morpheme and a word.

[2] In case the length of $\bar{e}_i$ is 1, e_1 is the same as $\bar{e}_i$.

3. Performance Enhancing Techniques

Performance evaluations are carried out on the C-STAR 2003 development test data consisting of 506 segments for both Japanese-to-English (J2E hereafter) and Chinese-to-English (C2E hereafter) translations. BLEU [9] has been used for translation quality evaluations, with 16 reference translations and the following evaluation parameters:

- Case insensitive
- Punctuations preserved

Translation model (TM hereafter) and language model (LM hereafter) training corpora are specified in Table 1.

Language	TM data	LM data
J2E	20K sentence pair BTEC[3]	380K word BTEC 140M word ViaVoice
C2E	20K sentence pair BTEC	380K word BTEC 4.9M word FBIS[4]

Table 1: Training corpora specifications

Across all evaluation conditions, both TM and LM are trained on lowercased English with punctuations preserved.

3.1. Baseline system

Baseline system performances are given in Table 2.

Languages	J2E	C2E
Baseline	**0.2924**	**0.2664**

Table 2: Baseline system performances

The key properties of the baseline system include (i) block selection from high precision word alignments using the projection and the extension algorithm, (ii) monotone decoding using block unigram probability, and word trigram language model probability at target phrase boundaries only.

3.2. Block selection from high recall alignment

We have found it effective to select blocks from high recall word alignments according to the projection algorithm and then filter out blocks which do not satisfy a length ratio between the source and the target phrase.

3.2.1. Chinese-to-English

Filter out blocks if they satisfy the condition (3):

(3) *target phrase length* ≥ *source phrase length * 2.5*

[3] **B**asic **T**raveler's **E**xpression **C**orpus distributed for the supplied data track training.

4 English-Chinese parallel corpus distributed by **F**oreign **B**roadcast **I**nformation **S**ervice.

Target and source phrase length ratio − *2.5* in (3) − is determined empirically. We start with a value higher than the source and target sentence length ratio (1.03 in our training corpus) and increase the value until the system finds the optimal value.

3.2.2. Japanese-to-English

Block selection for Japanese-to-English translation takes place in three steps: Step 1 − Morphological analysis as a preprocessing to TM training, [7]. Step 2 − Block selection from high recall word alignments & filtering according to the source and target length ratio. Step 3 − Merge blocks with the same source phrase to be translated into punctuations **.** and ***?***.

Morphological analysis: Japanese overtly marks the sentence types using sentence particles, as in (4) and (5):

(4) 革 見本 を みせ て いただけ ます **か** 。
Can you show me leather samples ?
(5) 毒虫 に 刺さ れ まし **た** 。
I was stung by a poisonous insect .

The question sentence (4) is marked by the particle か and the statement (5) by the particle た. As shown in (5), the role played by a sentence particle is often repeated by a punctuation (。). We delete sentence particles including う, ね, が, た, の, わ before TM training. The morphemes undergoing deletion analysis are typically those with a high null word translation probability.

Block selection and filtering: We obtain word alignments between English and morphologically analyzed Japanese parallel corpus. We apply the projection algorithm to high recall word alignments and filter out blocks satisfying the condition (6).

(6) *target phrase length > source phrase length * 1.5*

The value for the target and source phrase length ratio − *1.5* in (6) − is determined empirically in the manner described for C2E translation.

Merge blocks with fixed translations: We merge blocks containing source phrases to be translated into the question marker "?" and the period "." to insure that these source phrases are always correctly translated.

Performance improvement by block selection from high recall word alignments and filtering is shown in Table 3.

Languages	J2E	C2E
Baseline	0.2924	0.2664
Union + Filtering	**0.3249**	**0.2895**

Table 3. Impact of block selection from high recall word alignments and filtering

3.3. Reordering and block combination

Japanese and English word orders display a high degree of distortion primarily because the Japanese default word order is subject-object-verb whereas the English default word order is subject-verb-object, as in (7).

(7) ［ジャケット を］$_{object}$［探し て い ます］$_{verb}$ 。
I'm looking for a jacket.

We also observe word order discrepancies between Chinese and English questions, as indicated by the underlines in (8).

(8) 日本 航空 公司 的 柜台 **在 哪里** ？
Japan airline counter is where
Where is the Japan airline counter?

We identify words and phrases that indicate a *high degree of distortion* between the source and the target sentences, for example, by viterbi alignment. We then reorder the source language sentence into the target language word order, as in (9) and (10):

(9) ［ジャケット を］$_{object}$［探し て い ます］$_{verb}$ 。
→［探し て い ます］$_{verb}$［ジャケット を］$_{object}$ 。
(10) 日本 航空 公司 的 柜台 **在 哪里** ？
→ **哪里 在** 日本 航空 公司 的 柜台？

With reordering of source language sentences, we obtain two sets of parallel corpora: one in which no reordering is applied, and the other in which reordering is applied to the source language corpus. We acquire two sets of blocks from the two sets of parallel training corpora. We combine the two sets of blocks and recompute the block unigram probabilities.

Performance improvement by reordering and block combination is shown in Table 4.

Languages	J2E	C2E
Baseline	0.2924	0.2664
Union + Filtering	0.3249	0.2895
Reorder+Combine blocks	**0.3460**	**0.2957**

Table 4: Impact of reordering and block combination

We conjecture that the performance improvement by reordering and block combination is partially due to improvement in HMM word alignment. As pointed out in [8], HMM alignment is good at capturing local distortion whereas distortion models in the IBM source channel models are better at capturing long distance distortion. Reordering source language sentences into the target language word order results in either monotone alignment or local distortion between the source and the target languages.

3.4. Chinese unknown word segmentation

We derive a list of Chinese vocabulary and word bigrams from the word segmented Chinese training corpus. We apply unknown word segmentation as follows:

For each word *w* in the input, check to see if *w* occurs in the vocabulary list. If *w* does not occur in the vocabulary list, compute all possible segmentations of *w* at each character position. For example, if *w* consists of three characters $C_1C_2C_3$, then there are four possible segmentations.

Segmentation 1: $C_1C_2C_3$
Segmentation 2: C_1C_2 C_3
Segmentation 3: C_1 C_2C_3
Segmentation 4: C_1 C_2 C_3

For each segmentation, check to see if each two sub-word sequence occurs in the bigram list.

i. Select the segmentation with the least number of sub-words not covered by bigrams. Suppose Segmentation 2 and Segmentation 4 contain bigram sequences as shown below, where the italicized boldface indicates bigrams seen in the training corpus:

Segmentation 2: $\boldsymbol{C_1C_2}$ $\boldsymbol{C_3}$
Segmentation 4: $\boldsymbol{C_1}$ $\boldsymbol{C_2}$ C_3

All sub-words in Segmentation 2 are covered by bigrams, whereas C_3 is not covered by a bigram in Segmentation 4. Therefore, Segmentation 2 is selected.

ii. If more than one segmentation is equally covered by bigrams, select the segmentation with the least number of sub-words. Suppose Segmentation 2 and Segmentation 4 are covered by bigrams as shown below:

Segmentation 2: $\boldsymbol{C_1C_2}$ $\boldsymbol{C_3}$
Segmentation 4: $\boldsymbol{C_1}$ $\boldsymbol{C_2}$ $\boldsymbol{C_3}$

Segmentation 2 is chosen in this case since it contains two sub-words, whereas Segmentation 4 contains 3 sub-words.

iii. If more than one segmentation is equally covered by bigrams and contain the same number of sub-words, the segmentation with the most number of characters in the first sub-word is selected. Suppose Segmentation 2 and Segmentation 3, as shown below:

Segmentation 2: $\boldsymbol{C_1C_2}$ $\boldsymbol{C_3}$
Segmentation 3: $\boldsymbol{C_1}$ $\boldsymbol{C_2C_3}$

Since the first sub-word in Segmentation 2 contains 2 characters and the first sub-word in Segmentation 3 contains 1 character, Segmentation 2 is selected.[5]

Performance improvement by unknown word segmentation is shown in Table 5.

Languages	C2E
Baseline	0.2664
Union + Filtering	0.2895
Reorder + Combine phrases	0.2957
Unknownword segmentation	**0.3111**

Table 5: Impact of unknown word segmentation

3.5. Skip operation in decoding

We adopt *skip* operation for non-monotone decoding, [12], to capture the word order variations between the source and the target languages.

Skip is applied to delay translating one or more source phrases in case the current target phrase should be placed after subsequent target phrases to generate an accurate target sentence word order. We explain the intuition using the example (7):

(7) ［ジャケット を］$_{\text{object}}$［探し て い ます］$_{\text{verb}}$ 。
I′m looking for a jacket.

Suppose there are two blocks shown in (11) and (12), which cover the entire source word sequence.

(11) a jacket | ジャケット を
(12) I´m looking for | 探し て い ます

To decode the sentence (7), the system selects the blocks (11) and (12). If the system processes the blocks (11) and (12) monotonically, it will produce an inaccurate translation output "*a jacket I´m looking for* ." On the other hand, if the system *skips* to process the block (11) until after it processes (12) and translates 探し て い ます first, it will produce an accurate translation output "I´*m looking for a jacket*."

We impose two sets of constraints on skip, stated in (13) and (14), to prune out highly improbable word order sequences in advance.

(13) *Do not skip a block whose source phrase ends with a delimiter*.
(14) *Do not skip across a block whose source phrase starts with a delimiter*.

Delimiters are a set of punctuations and function words across which word orders do not change. Any source

[5] Conditions (*i*) to (*iii*) can be easily subsumed by incorporating language model probabilities derived from the training corpus, [10], [11], for language model based word segmentation.

word occurring to the left of a delimiter should occur to the left of the (translation of the) delimiter in its translation. Any source word occurring to the right of a delimiter should occur to the right of the (translation of the) delimiter in its translation. A set of delimiters we use include but not restricted to {. ? ，。、か 吗}. Delimiters can be automatically acquired by identifying the source words for which there is no crossing between the words to their left and the words to their right in viterbi alignment for each language pair.

Performance improvement by skip is shown in Table 6.[6]

Languages	J2E	C2E
Baseline	0.2924	0.2664
Union + Filtering	0.3249	0.2895
Reorder+ Combine blocks	0.3460	0.2957
Unknown word segmentation		0.3111
Skip in decoding	**0.4228**	**0.3470**

Table 6: Impact of skip operation in decoding

3.6. Language model probability computation

Instead of computing trigram language model probabilities only for words occurring at target phrase boundaries, we compute LM probabilities for each word in a target phrase, as ***schematically*** shown in (15).

(15) $p(\bar{e}_i \mid \bar{e}_{i-1}) = a * b * c$

a. First word of $\bar{e}_i$: $p(e_1 \mid e_h, e_{h-1})$

b. Second word of $\bar{e}_i$: $p(e_2 \mid e_1, e_h) * \alpha$

c. Subsequent words of $\bar{e}_i$: $p(e_j \mid e_{j-1}, e_{j-2}) * \alpha$

$\bar{e}_i$ is the current target phrase for which the LM probability is computed. $\bar{e}_{i-1}$ is the target word sequence in the translation hypothesis. e_h is the last word in $\bar{e}_{i-1}$. e_{h-1} is the second to the last word in $\bar{e}_{i-1}$. e_1 and e_2 are the first and the second word of $\bar{e}_i$, respectively. e_j is the j^{th} word in $\bar{e}_i$, where j > 2. LM score for e_1 and e_j (j > 1) may be differentiated by different weights denoted as α in (15b, c). The value for α may be parameterized for different language pairs.[7]

Once we compute the LM probability for each word in a target phrase, the system tends to generate less words than when we compute the LM probability for words at phrase boundaries only. We offset this side effect by adjusting the word generation penalty, [13], so that the system produces more words in the translation output without losing accuracy.

Performance improvement by the refined LM probability computation and word generation penalty is shown in Table 7.

Languages	J2E	C2E
Baseline	0.2924	0.2664
Union + Filtering	0.3249	0.2895
Reorder + Combine blocks	0.3460	0.2957
Unknown word segmentation		0.3111
Skip operation	0.4228	0.3470
LM+word generation penalty	**0.4307**	**0.3728**

Table 7. Impact of refined LM probability computation

3.7. Correlation between block selection and skip

While the experimental results in previous sections indicate that improvements in block selection and decoding techniques improve the translation quality independent of each other, there is an indication that the performance improvement by skip correlates with various block selection techniques.

Table 8 shows the apparent correlation in performance improvement by skip according to various block selection techniques for Japanese-to-English translation.

Block selection	Intersection +Extension	Union	Union+Filtering+ Reorder+Combine
Baseline decoding	0.2924	0.3100	0.3460
Skip	0.3181 +8.8%	0.3513 +13.3%	0.4228 +22.2%
Refined LM	0.3525 +10.8%	0.3763 +7.1%	0.4307 +1.9%

Table 8. Performance improvement by skip according to various block selection techniques

Block selection from intersection − high precision word alignment − according to the extension algorithm is used in our baseline system. Block selection from union − high recall word alignment − results in a performance improvement over the baseline (BLEU score improvement from 0.2924 to 0.3100). Combining blocks derived from reordered and un-reordered training corpora using high recall word alignment (union) and filtering results in the best performance with the baseline decoding (BLEU score 0.3460).

Performance improvement by skip is most significant with the block selection technique which results in the highest BLEU score, i.e. 22.2% improvement. We posit that a good block selection technique is more likely to generate blocks whose source phrases coincide with natural units for reordering, e.g. the object ジャケット を and the

[6] With skip operation, application of reordering to Japanese input does not yield a better performance than without reordering. The BLEU scores for Japanese-to-English translation in Tables 6 and 7 are obtained from the Japanese input without reordering.

[7] We have set α to 1 for Japanese-to-English and 1.21 for Chinese-to-English translation in the LM cost formula where the LM probability is represented as a cost using sum of –log likelihood.

verbs 探し て い ます in (7), accounting for the significant performance improvement by skip.

However, performance improvement by the refined LM probability is least significant with the block selection which results in the highest BLEU score, i.e. 1.9% improvement. We attribute this to an overlap in roles played by LM probability and other constraints on block selection. LM probability computation of each word in a target phrase is equivalent to filtering out some candidate blocks whose target phrase LM probabilities are less likely than others.

4. Spoken language translation evaluation

We address issues to be worked out before adopting an automatic evaluation metric as a single means of conversational speech translation evaluation: (i) characteristics of spoken language dialogs which typically do not occur in written texts, and yet significantly contribute to the information content of the entire utterance and (ii) lack of correlation between human and automatic evaluations. All examples in this section are taken from the BTEC training corpus.

4.1. Characteristics of spoken language dialogs

Speech Act: Spoken language dialogs crucially depend on speech acts for successful communications such as questions, requests, suggestions as well as statements, [17], [18], [20]. For instance, out of 20k segments in the BTEC training corpus for Japanese-to-English translation, 7,438 segments contain questions (denoted by the question marker *?*), and at least 1,775 segments contain requests (denoted by the phrase *please*). Examples are given in (16)−(19).

(16) To the zoo **?**
(17) This row empty **?**
(18) And the number and name of the person you are calling **?**
(19) A seat in the back, **please**.

The fact that (16)−(18) are questions – as opposed to a statement, as in "*I would like to go to the zoo.*" – can be construed only by the question mark "*?*".[8] The fact that (19) is a request (as opposed to a question, as in "*Do you have a seat in the back?*") can be construed by the function word "*please*".

Negation: Spoken language dialogs often center around the notion of affirmation/negation, especially if the utterances are expressed by yes–no questions. Out of 20k segments in the BTEC training corpus for Japanese-to-English translation, 329 segments contain some form of negation, as shown in (20)–(23).

(20) I ***ca n't*** have dessert, really .
(21) ***No***, I just got here .
(22) ***Do n't*** take too much off the top .
(23) I ***do n't*** quite understand .

Negation typically applies over the entire utterance, and incorrect translation of negation often leads to an interpretation opposite to what has been intended by the speaker.

Examples (16)–(19) suggest that punctuations play a major role in an accurate interpretation of speech act in conversational speech translation, and therefore should be included as a legitimate vocabulary in the evaluation. The real question is how much weight should be given to the information conveyed by speech act. Speaking in terms of BLEU, is it sufficient to treat speech act as one more vocabulary item and subsume it under the modified precision and brevity penalty, or we need a third parameter − speech act − in the scoring formula and assign an appropriate weight? (20)–(23) indicate that information conveyed by negation is more significant than that conveyed by other lexical items. Loss of negation in (23), as in "*I do quite understand*" is very likely to result in a communication failure, whereas loss of the adverb *quite*, as in "*I don't understand*" is not.

Given the significant role played by speech acts and negation, [19], it seems worthwhile to conduct experiments to precisely measure their impacts on overall translation quality and incorporate them to an automatic evaluation metric accordingly.

4.2. Correlation between human and automatic evaluations

Table 9 shows the ranks of our system (out of 9 systems) submitted to the Chinese-to-English unrestricted data track.

Evaluation Methods	Ranks
Human−Fluency	4
Human−Adequacy	**6**
BLEU	**3**
GTM	**2**
NIST	**3**
PER	4
WER	3

Table 9. Ranks of a system by various evaluation methods

Human-Fluency and Human-Adequacy indicate human evaluation of translation fluency and adequacy,

[8] Speech acts are often denoted by sentence particles in Japanese such as か for a question, う for a proposal , た for a statement, as well as a phrase ください for a request.

respectively. BLEU, GTM, NIST, PER and WER are 5 automatic evaluation metrics used in the evaluation.

Apparently, ***automatic evaluations and Human-Adequacy judgment do not correlate***, contrary to what has been reported in previous studies, [9] for BLEU, [14] for GTM, [15] for NIST, where they all report a strong correlation between automatic and human evaluations. The lack of correlation between human adequacy judgment and automatic evaluations might be attributed largely to two factors: One to different genre material and the other to different evaluation parameters.

Genre: The current evaluation focuses on spoken dialogs consisting of short sentences (8.7 words/sentence on average for Japanese and 7.6 words/sentence on average for Chinese) with many variations in dialog acts (e.g. question, statement, request, etc.), whereas the previous studies focus mainly on written news texts.

Evaluation Parameters: The current evaluation evaluates all lowercased translation output without any punctuations. In addition, part-of-speech tagging is applied to automatic evaluations but not to human evaluations.[9] However, previous studies – reporting a strong correlation between automatic and human evaluations – base their studies on translation output and reference translations where both punctuations and upper/lowercase distinctions are preserved. We have pointed out in Section 4.1 the potential significance of punctuations in conversational speech translation. [15] reports that upper/lower case distinction needs to be preserved in order for automatic evaluations to correlate with human evaluations. Furthermore, none of the previous studies have applied part-of-speech tagging in automatic evaluations.

Setting all evaluation parameters the same for the current evaluation as previous studies would shed light on the cause for the lack of correlation between human and automatic evaluations. If it turns out that human evaluations still do not correlate with automatic evaluations even after setting all evaluation parameters the same, it would indicate that conversational speech translations require a new evaluation metric to adequately capture the characteristics of spoken language dialogs not present in written texts.

4.3. Correlation across automatic evaluations

Table 10 shows some of our automatic evaluation scores of Chinese-to-English translations.

System	BLEU	GTM	NIST
C2E_1	**0.3619**	0.6819	7.3512
C2E_2	0.3464	0.6719	7.2893
C2E_3	0.3289	**0.6933**	**7.9626**

Table 10: Chinese-to-English automatic evaluation scores

BLEU and GTM/NIST scores do not correlate with each other. BLEU score is the highest for the system C2E_1, whereas GTM/NIST scores are the highest for the system C2E_3. Our experiments on C-STAR 2003 development test set show that BLEU score difference of about 0.03 is statistically significant at 95% confidence interval, indicating that the BLEU score difference of 0.033 between C2E_1 (0.3619) and C2E_3 (0.3289) is very likely to be statistically significant.[10]

With the caveat that the evaluation parameters are different for the current evaluations from previous studies, the scores in Table 10 suggest that some automatic evaluation metric should fit better for spoken language translation evaluation than others. Note that BLEU, GTM and NIST all incorporate the notion of *precision* and the *length ratio* between the translation output and the reference translation into their scoring formula. BLEU and GTM crucially differ in the way how length ratio is computed. *Brevity penalty* (BP) plays a less significant role than *precision* in BLEU whereas *recall* plays an equally important role as *precision* in GTM. Spoken language translation evaluation could serve as a test bed for differentiating the fitness of some version of length ratio to the overall translation evaluation task, which is not easily distinguishable in an evaluation of written news texts.[11]

5. Future Work

Recent success in machine translation of texts with the adoption of automatic evaluation metric BLEU indicates that a good evaluation metric correlating well with human judgments drives the machine translation technology development.

To come up with a good spoken language translation evaluation metric, however, there are at least two major issues to be worked out. First, we need to figure out what is the correct format of the reference translations to be used by human assessors. A good first approximation might be the format consistent with human transcriptions of speech. Second, we need to

[9] This information is due to personal communications with Michael Paul.

[10] [Paul et al. 2004] also show the lack of correlation between BLEU and NIST scores of the systems evaluated in C-STAR spoken language translation evaluation in 2003.

[11] According to [Melamed et al. 2003], BLEU and GTM both correlate well with human adequacy judgments on documents of more than 10 segments with more than 1 reference translation.

factor out characteristics of spoken language not present in written texts and decide whether or not these need to be introduced as independent parameters in the evaluation metric.

We believe that the notion of precision and brevity penalty in BLEU are applicable to all types of machine translation quality evaluations, and should serve as the baseline parameters for an improved spoken language translation evaluation metric which will drive a rapid improvement of the technology.

6. Acknowledgements

We would like to thank Yuqing Gao for the Chinese-English parallel corpus we have used in the Chinese-to-English unrestricted data track and Fei Xia for her Chinese word segmentation system.

7. References

[1] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer. "The mathematcs 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] Y. Al-Onaizan, Niyu Ge, Y-S. Lee, K. Papineni, "IBM Site Report", *Proceedings of DARPA Machichine Translation Evaluation Workshop*, Alexandria, VA, USA, June 22–23, 2004.

[22] F. Xia and M. McCord, "Improving a statistical MT system with automatically learned rewrite patterns", *Proceedings of COLING–2004*, pages 508–514, 2004.