Overview of the IWSLT 2005 Evaluation Campaign

Matthias Eck and Chiori Hori
Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{matteck, chiori}@cs.cmu.edu

Abstract
This paper reports an overview of the evaluation campaign results of the IWSLT 2005 workshop\(^1\). The BTEC corpus, which consists of typical travel domain phrases, was used. Data for the five language pairs Arabic/Chinese/Japanese/Korean to English and English to Chinese was prepared. To study how much the amount of the training data and how much different training and decoding approaches contribute to the performance, a supplied data and an unrestricted data track were introduced. In addition, translation results were evaluated not only for text input but also speech recognition output. 19 systems from 17 organizations participated in the evaluation. All machine translation results were evaluated using automatic evaluation metrics. The most popular track, translating text form Chinese to English, was graded by 3 humans in terms of Fluency, Adequacy and Meaning Maintenance. The correlation between automatic evaluation metrics and human judgment was examined.

1. Introduction
The Consortium for Speech Translation Advanced Research (C-STAR) had been formed in the 1990s to study and develop techniques for speech-to-speech translation. To further this research C-STAR members have been jointly constructing a multilingual spoken language corpus, the basic travel expression corpus (BTEC, [1]). In 2004 the International Workshop on Spoken Language Translation (IWSLT) was started in order to enable the exchange of knowledge among researchers working on speech-to-speech translations and to create an opportunity to enhance the machine translation (MT) systems by comparing technologies on the same test bed [2]. IWSLT 2005 extended over the 2004 evaluation campaign by translating the output of automatic speech recognition (ASR) systems as well.

Speech-to-speech translation systems are designed as systems combining an MT system with automatic speech recognition technology. This introduces additional difficulties into the translation process, caused not only by disfluencies due to the spontaneity of the spoken language but also due to the errors in the ASR output. To accomplish speech-to-speech translation, we have to solve problems in translating spoken language and handling ASR output.

This was the reason why we focused on handling speech recognition output including multiple recognition hypotheses this year. Translating speech with the goal of maintaining the original information in the source speech makes it necessary to handle recognition errors. Some of the problems that will have to be addressed are:

- How can ASR output be translated more accurately even if recognition errors exist?
- How much could the MT performance be enhanced by considering multiple hypotheses?
- Which hypothesis contributes the most to the MT performance?
- How to select the best hypothesis that can be translated well from multiple hypotheses?

To alleviate the difficulty to work on both speech recognition and machine translation, we provided speech recognition results in this workshop. Future evaluations might also include a speech recognition part evaluating the attendees’ ASR systems. More realistic and difficult data such as spontaneous conversational speech could also be used.

One outcome of the evaluation campaign is that a large number of the simple BTEC sentences can already be correctly translated but there are still open questions, especially when translating BTEC sentences already be translated but there are still open questions, especially when translating BTEC sentences

Finally, we hope that IWSLT will continue to provide opportunities to compare the technologies and give answers to scientific questions addressed in the field of speech-to-speech translation.

\(^1\) Workshop website:
http://www.is.cs.cmu.edu/iwslt2005/
2. Evaluation Conditions

2.1. Language pairs and source input

The language pairs that were used for the IWSLT 2005 evaluation are shown in Table 1. Chinese, Japanese, Arabic and Korean were translated to English. English was translated to Chinese. Manual transcriptions were provided for all tracks as source input. Speech recognizer output (ASR output) was provided as source input for Chinese, Japanese and English in the form of n-best lists and word lattices.

| Translation direction | Manual transcription | ASR output |
|-----------------------|----------------------|------------|
| Chinese → English     | ✓                     | ✓          |
| Japanese → English    | ✓                     |            |
| Arabic → English      | ✓                     | ✓          |
| Korean → English      | ✓                     |            |
| English → Chinese     | ✓                     | ✓          |

Table 1: Translation directions and input type

2.2. Data Track Conditions

Four different data track conditions were distinguished.

- In the “Supplied” data track the training data was limited to 20,000 sentence pairs with given word segmentation for Chinese, Japanese, and Korean.
- The “Supplied Data & Tools”-track permitted the additional use of Natural Language Processing (NLP) Tools like Taggers, Chunkers, or Parsing Tools (which could be trained on additional data).
- The “Unrestricted” data track allowed the use of any publicly available data besides the supplied data and the NLP-Tools. Mainly LDC resources or web data could be applicable here.
- The “C-STAR”-track even allowed the use of proprietary data. In this case the whole BTEC corpus, which is available to the members of the C-STAR consortium, was the most significant proprietary data, as this is then additional in-domain data.

The data track limitations applied to both bilingual as well as monolingual corpora for translation model and language model training respectively. Table 2 gives an overview of the permitted (✓) and not permitted (X) linguistic resources.

2.3. Characteristics of Training Data & Test Sets

All supplied training data, as well as the development and test sets were taken from the BTEC corpus, which consists of typical phrases and sentences from the travel domain [1]. Table 3 shows some example sentences (from the IWSLT 2005 test set).

| Lines | Words | Vocabulary |
|-------|-------|------------|
| Chinese | 176,199 | 8,687 |
| Japanese | 198,453 | 9,277 |
| Arabic  | 131,711 | 26,116 |
| Korean  | 208,763 | 9,132 |
| English | 183,452 | 6,956 |

Table 4: Characteristics of the IWSLT 2005 supplied training data
Table 5: Characteristics of the IWSLT 2005 test sets (for manual transcription)

| Language (→ Translation) | Lines | Words | Vocabulary | Unknown words |
|--------------------------|-------|-------|------------|--------------|
| Chinese (→ English)     | 506   | 3,743 | 963        | 155          |
| Japanese (→ English)    |       | 4,226 | 975        | 169          |
| Arabic (→ English)      |       | 2,607 | 1,387      | 468          |
| Korean (→ English)      |       | 4,563 | 969        | 84           |
| English (→ Chinese)     |       | 3,897 | 842        | 66           |

**ASR output**

The ASR outputs were provided by ATR, NLPR and UKA. Lattices and n-best lists for Chinese, Japanese and English were generated. Table 6 lists the word and sentence error rates for the Chinese (Mandarin) and Japanese n-best list. Overall the Japanese error rates are much lower than the error rates for Chinese. As no-one submitted translations for ASR output, English to Chinese, we did not include the error rates here.

Table 6: Characteristics of the provided ASR output

| Language (→ Translation) | 1-best word error rate | 1-best sentence error rate | 20-best word error rate | 20-best sentence error rate | best sentence in 20-best (avg. rank) |
|--------------------------|------------------------|---------------------------|-------------------------|----------------------------|-------------------------------------|
| Chinese (→ English)     | 0.349                  | 0.715                     | 0.226                   | 0.510                      | 3.62                                |
| DEVSET1                  | 0.330                  | 0.724                     | 0.211                   | 0.492                      | 3.35                                |
| DEVSET2                  | 0.349                  | 0.715                     | 0.226                   | 0.510                      | 3.62                                |
| IWSLT 2005               | 0.420                  | 0.806                     | 0.253                   | 0.557                      | 4.57                                |
| Japanese (→ English)    | 0.055                  | 0.196                     | 0.037                   | 0.132                      | 1.31                                |
| DEVSET1                  | 0.056                  | 0.198                     | 0.031                   | 0.132                      | 1.33                                |
| DEVSET2                  | 0.056                  | 0.198                     | 0.031                   | 0.132                      | 1.33                                |
| IWSLT 2005               | 0.160                  | 0.603                     | 0.123                   | 0.528                      | 2.06                                |

Table 7: English Evaluation specification

**English – Standard Evaluation**
- case insensitive, all lowercase
- removed punctuation marks \'. ' '"', '"','""','"""','""""'"
- removed '-' to split compounds

**English – Optional Evaluation**
- case sensitive, mixed case
- separated punctuation marks

Table 8: Chinese Evaluation specification

2.4. Evaluation Specifications

The main objective of IWSLT 2005 is the evaluation of speech-to-speech translation techniques. Since we consider the punctuation marks and mixed casing to be less relevant here, the standard evaluation does not distinguish between upper/lowercase letters and does not consider punctuation marks. This applied to all English outputs; all translations being automatically preprocessed by the evaluation server.

To evaluate true-casing techniques and the preservation or insertion of punctuation marks an optional evaluation was offered. The optional evaluation was done case sensitive with separated punctuation marks. This evaluation was only done if the submitted translation contained uppercase characters. The punctuation marks in this case were automatically separated.

Table 7 gives an overview of the English evaluation specifications.

Table 9 gives short descriptions of the used automatic evaluation metrics.

1 Not all 16 references were created as genuine translations but as paraphrases.
Generally automatic metrics compare the translations with references manually generated by humans. The metrics are usually based on edit distance or n-gram precision. The problem here is that the number of references being limited, all possible references are not covered. Consequently, even a perfect translation can not be correctly evaluated if the appropriate reference is missing. To alleviate such influence of insufficient references, some metrics count “quasi references” by combing parts of phrases in all references. Though this can help, there is no guarantee that it simulates the performance we would have with all possible references. The ideal condition would be to prepare the nearest references for each translation result. But this would still be expensive.

### Table 9: Automatic Evaluation metrics

| Metric | Description |
|--------|-------------|
| BLEU   | The geometric mean of n-gram precision by the system output with respect to reference translations. Scores range between 0 (worst) and 1 (best). |
| NIST   | A variant of BLEU using the arithmetic mean of weighted n-gram precision values. Scores are positive numbers with 0 being the worst possible score. |
| mWER   | Word Error Rate on multiple references: The edit distance between the system output and the closest reference translation. Scores range between 0 (best) and 1 (worst). |
| mPER   | Position independent mWER: a variant of mWER that disregards word ordering. Scores range between 0 (best) and 1 (worst). |
| GTM    | Measures the similarity between texts by using a unigram-based F-measure. Scores range between 0 (worst) and 1 (best). |
| METEOR | Scoring method that matches translations with the references in different stages. Exact matches, stem matches and synonym matches are considered. Meteor does not distinguish between lower and upper case. It is not yet able to reliably score Chinese output. Scores range between 0 (worst) and 1 (best). |

### Table 10: Adequacy and Fluency judgments

| Fluency  | Adequacy |
|----------|----------|
| 0 Incomprehensible | 0 None |
| 1 Disfluent English | 1 Little information |
| 2 Non-Native English | 2 Much information |
| 3 Good English | 3 Most information |
| 4 Flawless English | 4 All information |

### Meaning Maintenance

We also tried another metric in order to extend the Adequacy scoring. The metric Meaning Maintenance intends to compare the meaning of the translation with the source. This metric is more concerned with the actual meaning of a translation. If a translation error is rather obvious and does not change the general meaning, the translation will still be useful. If the meaning is completely twisted, for example negated, the translation will not be useful and this mistake has to be avoided. In Adequacy judgments human graders might tend to ignore this misleading information and grade only the correct parts. The Meaning Maintenance score tries to distinguish between degrees of additional information in the translations. If information that was added during the translation was misleading the Meaning Maintenance score would be very low even if the rest of the translation was correct. The Adequacy score would probably not be as low. It is however obvious that there will be a high correlation between Adequacy and Meaning Maintenance scores. Table 11 shows the different scores assigned for Meaning Maintenance and short explanations.

### Table 11: Meaning Maintenance

| Meaning Maintenance | Description |
|---------------------|-------------|
| 0                   | Totally different meaning |
| 1                   | Partially the same meaning but misleading information is introduced |
| 2                   | Partially the same meaning and no new information |
| 3                   | Almost the same meaning |
| 4                   | Exactly the same meaning |

Fortunately, we have 16 references per sentence for translations to English. Although it is difficult to say if 16 references can cover all possible references, we can examine various phenomena using many more references in comparison with other test beds.

### 2.6. Subjective Evaluation

The subjective evaluation was done on the most popular track, Chinese to English, translation of manual transcriptions with supplied data. 11 systems were submitted to this track and all of them were evaluated by bilingual human graders. At the time of the paper deadline 3 graders had finished the evaluation. Every grader evaluated 10% of the sentences twice to check for inconsistencies.

**Adequacy and Fluency**

The typically used metrics for subjective evaluation are Fluency and Adequacy [10]. Fluency corresponds to the degree to which the translation is well-formed as per the target language, disregarding the meaning of the original source sentence. Adequacy refers to the degree to which the translation preserves the original information present in the source sentence.

| Fluency  | Adequacy |
|----------|----------|
| 0 Incomprehensible | 0 None |
| 1 Disfluent English | 1 Little information |
| 2 Non-Native English | 2 Much information |
| 3 Good English | 3 Most information |
| 4 Flawless English | 4 All information |

### Table 10: Adequacy and Fluency judgments
Subjective Evaluation procedure

To keep consistency in grading, all systems were displayed at the same time and evaluated by comparing all translations for one sentence. The translations were randomly ordered to avoid influences on the judgment by the position in the list. One of the reference translations was included among the machine translations to calibrate the translation quality and give an upper bound. First all Fluency scores were assigned for all sentences and all translations, then the Adequacy and finally the Meaning Maintenance scores.

For Fluency judgments the graders did not see the source sentence. Adequacy and Meaning Maintenance scores were evaluated by comparing the source sentence and translations results. This avoids any bias from the reference translations. To evaluate under the same condition used in the automatic evaluation all translations were preprocessed according to the standard evaluation by lower casing them and removing the punctuation marks. If there were any meaning uncertainties caused by the missing punctuation marks, the graders were asked to judge in favor of the system. Any remaining Chinese characters were deleted.

3. Participants and submissions

Sixteen groups (17 organizations, 2 organizations cooperated in one group) actively participated in the evaluation campaign of IWSLT 2005. Table 18 in Appendix A lists all participants and the techniques that were used by each of the systems. Since one of the institutions submitted three systems (ATR) and a second institution submitted two systems (TALP), 19 systems were submitted in all.

By far the majority of systems were statistical machine translation systems (SMT), some of which used additional syntax information. Three systems used the example based machine translation (EBMT) technique and one system used the output of different translation engines as a multi engine system (MEMT).

Table 12 and Table 13 indicate the number of participants for each track. Sixty-nine translations were done using manual transcriptions as an input, 15 using ASR output in the form of lattices or n-best lists. A majority of 11 systems was submitted to the Supplied Data track translating manual transcriptions from Chinese to English. Submissions containing mixed case were rather rare with only 18 instances. There were no submissions for the translation of ASR output translating English to Chinese.

4. Evaluation Results

In this section we will investigate some general tendencies and overall results. For all detailed evaluation scores please refer to Appendix B. Section B.1 shows the results of the subjective evaluation compared with the automatic scores. The latter sections list the automatic scores for all translation directions, tracks and data conditions.

4.1. Analysis of the Automatic Evaluation

All of the automatic metrics focus on different features to define their scores. Because of that it is not surprising that they rank systems differently. The subjective impression is however that the general tendency (good translation – bad translation) stays the same for all metrics. This is supported by the Pearson correlations of the automatic metrics shown in Table 14.
Table 14: Pearson correlation between automatic scores

|       | BLEU | NIST | mWER  | mPER  | GTM   | METEOR |
|-------|------|------|-------|-------|-------|--------|
| BLEU  | 1.00 | 0.77 | -0.97 | -0.94 | 0.85  | 0.82   |
| NIST  | 1.00 | -0.74| -0.85 | 0.72  | 0.77  |        |
| mWER  | 1.00 | 0.97 | -0.90 | -0.74 |       |        |
| mPER  | 1.00 | -0.91| -0.81 |      |       |        |
| GTM   |      |      | 1.00  | 0.64  |       |        |
| METEOR|      |      |       | 1.00  |       |        |

The highest correlation was observed between mWER and the BLEU score and mPER and mWER; the lowest correlation between GTM and METEOR. Figure 1 illustrates the correlation between the mainly used NIST and BLEU scores in a diagram.

Table 15: Correlation between automatic and subjective metrics (incl. 95% confidence intervals)

|       | Adequacy | Fluency | Mean. Maint. |
|-------|----------|---------|--------------|
| BLEU  | 0.70 [0.19, 0.92] | 0.95 [0.81, 0.99] | 0.75 [0.27, 0.93] |
| NIST  | 0.90 [0.67, 0.98] | 0.46 [0.17, 0.84] | 0.86 [0.53, 0.96] |
| mWER  | -0.72 [0.92, -0.22] | -0.90 [0.97, -0.66] | -0.79 [0.94, -0.35] |
| mPER  | -0.90 [0.97, -0.64] | -0.83 [0.95, -0.46] | -0.93 [0.98, -0.75] |
| GTM   | 0.89 [0.62, 0.97] | 0.74 [0.26, 0.93] | 0.93 [0.74, 0.98] |
| METEOR| 0.98 [0.92, 0.99] | 0.57 [0.04, 0.87] | 0.97 [0.89, 0.99] |

Grader inconsistency

10% of the sentences were evaluated twice by each grader to measure grader inconsistencies. The average differences between the first and second grade are listed in Table 16. (These inconsistencies were not considered for the confidence intervals in Appendix B.1, which were just calculated using standard statistical methods.)

Table 16: Average difference between first and second grade

|       | Adequacy | Fluency | Mean. Maint. |
|-------|----------|---------|--------------|
| Grader 1 | 0.32 | 0.29 | 0.25 |
| Grader 2 | 0.32 | 0.30 | 0.24 |
| Grader 3 | 0.60 | 0.61 | 0.40 |
| Average | 0.41 | 0.40 | 0.30 |

We can see that the average differences for Fluency and Adequacy are very similar at 0.40 and 0.41 which corresponds to the average differences reported for IWSLT 2004 [2]. The newly introduced Meaning Maintenance score however has only an average difference of 0.30. This indicates that a consistent grading is easier with the Meaning Maintenance score as the focus on “meaning” and the instructions give the grader a clear way to distinguish between the different grades.

Do we need Meaning Maintenance?

On the other hand, the overall scores indicate that Meaning Maintenance has a high correlation with Adequacy (Pearson: 0.82). Also, in 91% of the graded sentences the difference between Adequacy and Meaning Maintenance is less than 2.

It is however obvious that there will be little difference in the Adequacy and Meaning Maintenance scoring if the translation is very good and gets high scores. Therefore we investigated the correlation for sentences that got low scores (Meaning Maintenance 0 or 1). The average scoring difference here is 0.75 with a Pearson correlation of 0.20. The average scoring difference for high scores (Meaning Maintenance 3 or 4) is only 0.25 (Pearson 0.65 on the same number of
samples). This means that for good translations the graders tended to use very similar scores for Meaning Maintenance and for Adequacy but their scores differed for worse translations. However, the variation is generally higher for lower scores for all metrics which could also explain the above differences.

But we could show that consistent grading is easier with the Meaning Maintenance score. A reason could be that the focus on “meaning” and the instructions give the grader a clear way to distinguish between the different grades. Therefore, it could generally be valuable to use this metric in the future especially for more complicated translation tasks, for example the translation of news texts. A longer sentence could be much more twisted and additional information can be more misleading.

It will most probably not be necessary to introduce Meaning Maintenance as an additional score but it will be sufficient to change the instructions for Adequacy to make graders aware of misleading information. This will also help graders to score translations more consistently.

4.3. Example translations

A number of reference sentences with example translations taken from different submissions, different tracks, and data conditions are listed in Table 17. Some translations are completely perfect while others introduce additional misleading information.

| Reference: | are there any shops which sell reasonably priced bags |
|------------|-------------------------------------------------------|
| Translation: | are there any shops which sell reasonably priced bags |
| Translation: | are there any of my bag at reasonable price can I buy |
| Translation: | is the stationery store bags can I have a reasonable price |
| Translation: | there are some store |

| Reference: | i would like to have an allergy test please |
|------------|------------------------------------------------|
| Translation: | i would like to have an allergy test please |
| Translation: | could you check i am allergic |
| Translation: | i would like to make a |
| Translation: | allergic to order room service please |

| Reference: | i would like a room facing the beach |
|------------|-------------------------------------|
| Translation: | i would like a room facing the beach |
| Translation: | i would like a room that faced a beach |
| Translation: | i would like to the beach room |
| Translation: | i would like a in my room |

Table 17: Example translations

5. Acknowledgements

We thank Almut Silja Hildebrand for her extensive technical support with the Evaluation tools. Many thanks to Michael Paul for supporting the organization with his experience from last year’s IWSLT2004. Dr. Hirofumi Yamamoto, Dr. Zhenbiao Chen and Matthias Paulik supported the evaluation by generating the ASR outputs.

We also thank all institutions for participating in our evaluation campaign and for making IWSLT 2005 a success.
6. References

[1] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World. Proceedings of LREC 2002, Las Palmas, Spain.

[2] Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, Jun’ichi Tsuji. Overview of the IWSLT’04 Evaluation Campaign. Proceedings of IWSLT 2004, Kyoto, Japan.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of ACL 2002, Philadelphia, PA, USA.

[4] George Doddington, 2001. Automatic Evaluation of Machine Translation Quality using n-Gram Co-occurrence Statistics. NIST Washington, DC, USA.

[5] Sonja Niestlen, Franz Josef Och, Gregor Leusch, Hermann Ney. An evaluation tool for machine translation: Fast evaluation for machine translation research. Proceedings of LREC2000, Athens, Greece.

[6] Franz Josef Och. Minimum error rate training in statistical machine translation. Proceedings of ACL2003, Sapporo, Japan.

[7] Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of Machine Translation and its Evaluation. Proceedings of MT Summit IX, New Orleans, LA, USA, 2003.

[8] Satanjeev Banerjee, Alon Lavev. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. ACL’05 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA.

[9] Ying Zhang and Stephan Vogel. Measuring Confidence Intervals for the Machine Translation Evaluation Metrics. Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2004), Baltimore, MD, USA.

[10] John S. White, Theresa A. O’Connell, Francis E. O’Mara. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. Proceedings of AMTA 1994, Columbia, MD, USA.

[11] Richard Zens, Oliver Bender, Sasa Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuji Zhang, Hermann Ney. The RWTH Phrase-based Statistical Machine Translation System. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[12] Sanjika Hewavitharana, Bing Zhao, Almut Silja Hildebrand, Matthias Eck, Chiori Hori, Stephan Vogel and Alex Waibel. The CMU Statistical Machine Translation System for IWSLT 2005. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[13] Philipp Koehn, Amitai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, David Talbot. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[14] Kazuteru Ohashi and Kazuhide Yamamoto and Kuniko Saito and Masaaki Nagata. NUT-NTT Statistical Machine Translation System for IWSLT 2005. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[15] Steve DeNeefe, Kevin Knight. ISL’s 2005 Statistical Machine Translation Entries. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[16] Sadao Kurohashi, Toshiaki Nakazawa, Kauffman Alexis, Daisuke Kawahara. Example-based Machine Translation Pursuing Fully Structural NLP. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[17] Michael Paul, Takao Doi, Youngsook Hwang, Kenji Imamura, Hideo Okuma, Eiichiro Sumita. Nobody is Perfect: ATR’s Hybrid Approach to Spoken Language Translation. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[18] Yves Lepage and Etienne Denoual. ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[19] Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto. Using Multiple Recognition Hypotheses to Improve Speech Translation. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[20] Bo Xu, Chengqing Zong. The CASIA Phrase-Based Machine Translation System. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[21] Wei Pang, Zhendong Yang, Zhenbiao Chen, Wei Wei, Bo Xu, Yves Lepage and Etienne Denoual. The MIT-LL/AFRL MT System. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[22] Josep M. Crego, Adrià de Gisbert and José B. Mariño. The TALP Ngram-based SMT System for IWSLT-05. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[23] Hajime Tsukada, Taro Watanabe, Jun Suzuki, Hitomi Kazawa, and Hideki Isokazi. The NTT Statistical Machine Translation System for IWSLT 2005, Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[24] Marta R. Costa-Jussà and José A. R. Fonollosa. Tuning a phrase-based statistical translation system for the IWSLT 2005 Chinese to English and Arabic to English tasks. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[25] Young-Suk Lee. IBM Statistical Machine Translation for Spoken Languages. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[26] Arul Menezes and Chris Quirk. Microsoft Research Treetlet Translation System: IWSLT Evaluation. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[27] Miki Sasaki and Toshiki Murata. A Pattern-Based Machine Translation System — Yakushite Net MT Engine. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.

[28] Yookyung Kim, Jun Huang, Youssef Billawala, Demetrios Master, Farzad Ehsani. Sehda S2MT: Incorporation of Syntax into Statistical Translation System. Proceedings of IWSLT 2005, Pittsburgh, PA, USA, October 2005.
### Appendix A  System overview

| Institution                                      | Description in paper                                                                 | Technique | Short form for systems |
|-------------------------------------------------|--------------------------------------------------------------------------------------|-----------|------------------------|
| RWTH Aachen University                          | Zens et al. The RWTH Phrase-based Statistical Machine Translation System. [11]       | SMT       | RWTH                   |
| Carnegie Mellon University                      | Hewavitharana et al. The CMU Statistical Machine Translation System for IWSLT 2005. [12] | SMT       | CMU                    |
| University of Edinburgh                         | Koehn et al. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. [13] | SMT       | EDINBURGH              |
| Nagaoka University of Technology                | Ohashi et al. NUT-NTT Statistical Machine Translation System for IWSLT 2005. [14]     | SMT       | NGKUT                  |
| University of Southern California – Information Sciences Institute | DeNeefe and Knight. ISi's 2005 Statistical Machine Translation Entries. [15]               | SMT (Syntax) | USC-ISI |
| University of Tokyo                             | Kurohashi et al. Example-based Machine Translation Pursuing Fully Structural NLP. [16]    | EBMT      | UTOKYO                 |
| ATR Spoken Language Communication Research Labs | Paul et al. Nobody is Perfect: ATR's Hybrid Approach to Spoken Language Translation. [17] | MEMT      | ATR-C3                 |
|                                                 | Lepage and Denoaul. ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages. [18] | EBMT      | ATR-ALEPH              |
|                                                 | Zhang et al. Using Multiple Recognition Hypotheses to Improve Speech Translation. [19]   | SMT       | ATR-SLR                |
| ITC - Center for Scientific and Technological Research | Chen et al. The ITC-irst SMT System for IWSLT-2005. [20]                            | SMT       | ITC-IRST               |
| MIT/Lincoln Laboratory – Airforce Research Laboratory | Shen et al. The MIT-LL/AFRL MT System. [21]                                            | SMT       | MIT-LL/AFRL            |
| National Laboratory of Pattern Recognition       | Pang et al. The CASIA Phrase-Based Machine Translation System. [22]                    | SMT       | NLPR                   |
| NTT Cyber Space Laboratories                     | Tsukada et al. The NTT Statistical Machine Translation System for IWSLT 2005. [23]     | SMT       | NTT                    |
| TALP Research Center                            | Crego et al. The TALP Ngram-based SMT System for IWSLT'05. [24]                      | SMT       | TALP-ngram             |
|                                                 | Costa-jussà and Fonollosa. Tuning a phrase-based statistical translation system for the IWSLT 2005 Chinese to English and Arabic to English tasks. [25] | SMT       | TALP-phrase             |
| IBM Research                                    | Lee. IBM Statistical Machine Translation for Spoken Languages [26].                   | SMT       | IBM                    |
| Microsoft Research                              | Menezes and Quirk. Microsoft Research Treelet Translation System: IWSLT Evaluation. [27] | SMT (Syntax) | MICROSOFT              |
| Oki Electric Industry Co., Ltd.                 | Sasaki and Murata. A Pattern-Based Machine Translation System — Yakushite Net MT Engine. [28] | EBMT      | OKI                    |
| Sehda Inc.                                      | Kim et al. Sehda S2MT: Incorporation of Syntax into Statistical Translation System. [29] | SMT (Syntax) | SEHDA                  |

Table 18: Overview of the participating institutions and translation systems
# Appendix B  Evaluation Results

## B.1  Translation of manual transcription Chinese to English – Human Evaluation

| Human Evaluation | Fluency | Adequacy | Meaning Maintenance |
|------------------|---------|----------|---------------------|
| ITC-IRST         | 3.15    | [3.09, 3.21] | MIT-LL/AFRL 2.71 [2.64, 2.79] | MIT-LL/AFRL 2.63 [2.55, 2.70] |
| RWTH             | 3.04    | [2.98, 3.11] | ITC-IRST 2.65 [2.57, 2.72] | RWTH 2.60 [2.53, 2.68] |
| CMU              | 2.88    | [2.81, 2.94] | RWTH 2.63 [2.55, 2.70] | ITC-IRST 2.60 [2.53, 2.68] |
| ATR-C3           | 2.86    | [2.79, 2.93] | TALP-phrase 2.52 [2.45, 2.60] | TALP-phrase 2.49 [2.41, 2.56] |
| TALP-ngram       | 2.82    | [2.75, 2.86] | IBM 2.51 [2.44, 2.59] | IBM 2.44 [2.37, 2.52] |
| EDINBURGH        | 2.81    | [2.74, 2.87] | TALP-ngram 2.44 [2.37, 2.52] | TALP-ngram 2.40 [2.32, 2.47] |
| MIT-LL/AFRL      | 2.79    | [2.72, 2.85] | EDINBURGH 2.33 [2.25, 2.40] | EDINBURGH 2.35 [2.27, 2.43] |
| TALP-phrase      | 2.78    | [2.71, 2.84] | ATR-C3 2.31 [2.23, 2.39] | ATR-C3 2.23 [2.15, 2.31] |
| IBM              | 2.77    | [2.71, 2.84] | NTT 2.09 [2.02, 2.16] | NTT 2.03 [1.95, 2.10] |
| USC-ISI          | 2.32    | [2.25, 2.39] | CMU 1.95 [1.87, 2.03] | USC-ISI 1.96 [1.88, 2.03] |
| NTT              | 1.97    | [1.90, 2.04] | USC-ISI 1.90 [1.82, 1.97] | CMU 1.94 [1.86, 2.02] |

Human Evaluation scores for first reference translation:

- Fluency: 3.72 [3.68, 3.75]
- Adequacy: 3.68 [3.64, 3.73]
- Meaning Maintenance: 3.64 [3.59, 3.68]
## B.2 Translation of manual transcription Chinese to English – Automatic Evaluation

### Supplied Data

#### Standard Evaluation

| System         | BLEU Score | NIST Score |
|----------------|------------|------------|
| ITC-IRST       | 0.528      | [9.10, 9.99] |
| RWTH           | 0.511      | [9.95, 9.66] |
| EDINBURGH      | 0.465      | [8.60, 9.54] |
| TALP-phrase    | 0.452      | [8.02, 8.88] |
| MIT-LL/AFRL    | 0.450      | [7.93, 8.91] |
| CMU            | 0.444      | [7.44, 8.47] |
| IBM            | 0.440      | [7.15, 7.84] |
| ATR-C3         | 0.394      | [5.86, 7.05] |
| USC-ISI        | 0.332      | [5.48, 6.84] |
| NTT            | 0.278      | [5.01, 6.11] |

#### Mixed Case Evaluation

| System         | BLEU Score | NIST Score |
|----------------|------------|------------|
| ITC-IRST       | 0.528      | [8.30, 9.08] |
| IBM            | 0.450      | [7.67, 8.39] |

| System         | mWER       | mPER       | GTM       | METEOR     |
|----------------|------------|------------|-----------|------------|
| ITC-IRST       | 0.414      | 0.346      | 0.620     | 0.709      |
| RWTH           | 0.428      | 0.355      | 0.619     | 0.689      |
| EDINBURGH      | 0.453      | 0.358      | 0.609     | 0.665      |
| TALP-phrase    | 0.459      | 0.380      | 0.601     | 0.663      |
| MIT-LL/AFRL    | 0.464      | 0.391      | 0.599     | 0.652      |
| CMU            | 0.469      | 0.390      | 0.598     | 0.642      |
| TALP-ngram     | 0.482      | 0.408      | 0.567     | 0.632      |
| ATR-C3         | 0.513      | 0.428      | 0.553     | 0.629      |
| USC-ISI        | 0.544      | 0.469      | 0.524     | 0.567      |
| NTT            | 0.653      | 0.521      | 0.492     | 0.564      |
### Supplied Data + Tools

#### Standard Evaluation

| BLEU Score | NIST Score |
|------------|------------|
| IBM        | 0.479      | NGKUT      | 8.52 | [8.13, 8.91] |
| NGKUT      | 0.390      | USC-ISI    | 7.98 | [7.57, 8.37] |
| ATR-C3     | 0.380      | IBM        | 7.88 | [7.31, 8.40] |
| USC-ISI    | 0.376      | ATR-SLR    | 7.20 | [6.85, 7.56] |
| ATR-SLR    | 0.305      | ATR-C3     | 6.75 | [6.29, 7.26] |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| IBM  | 0.445 | IBM | 0.379 | IBM | 0.597 | NGKUT | 0.679 |
| USC-ISI | 0.537 | USC-ISI | 0.411 | USC-ISI | 0.576 | IBM | 0.651 |
| NGKUT | 0.538 | NGKUT | 0.419 | NGKUT | 0.568 | USC-ISI | 0.634 |
| ATR-C3 | 0.544 | ATR-C3 | 0.462 | ATR-C3 | 0.495 | ATR-C3 | 0.582 |
| ATR-SLR | 0.607 | ATR-SLR | 0.494 | ATR-SLR | 0.471 | ATR-SLR | 0.574 |

#### Mixed Case Evaluation

| BLEU Score | NIST Score |
|------------|------------|
| IBM        | 0.486      | IBM      | 0.367 | CMU | 6.92 | [7.24, 8.17] |
| NGKUT      | 0.292      | (0.267, 0.316) | NGKUT | 6.68 | [6.41, 6.95] |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| IBM  | 0.399 | IBM | 0.351 | IBM | 0.627 | IBM | 0.651 |
| NGKUT | 0.616 | NGKUT | 0.487 | NGKUT | 0.508 | NGKUT | 0.679 |

#### Unrestricted Data

#### Standard Evaluation

| BLEU Score | NIST Score |
|------------|------------|
| IBM        | 0.499      | CMU        | 9.35 | [8.90, 9.75] |
| CMU        | 0.471      | IBM        | 8.17 | [7.59, 8.73] |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| IBM  | 0.434 | CMU | 0.365 | CMU | 0.611 | CMU | 0.670 |
| CMU  | 0.469 | IBM | 0.372 | IBM | 0.610 | IBM | 0.663 |

#### Mixed Case Evaluation

| BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|------------|------------|------|------|-----|--------|
| IBM        | 0.500      | [0.467, 0.535] | 7.93 | [7.51, 8.41] | 0.387 | 0.345 | 0.639 | 0.662 |
## C-STAR Data

### Standard Evaluation

|               | BLEU Score | NIST Score |               |               |               |               |
|---------------|------------|------------|---------------|---------------|---------------|---------------|
| NLPR          | 0.528      | [0.496, 0.560] | NLPR          | 10.25         | [9.89, 10.61] |
| CMU           | 0.527      | [0.489, 0.563] | CMU           | 10.02         | [9.59, 10.43] |
| ATR-C3        | 0.503      | [0.462, 0.545] | ATR-C3        | 8.69          | [8.17, 9.17]  |
| ATR-ALEPH     | 0.477      | [0.439, 0.515] | ATR-SLR       | 8.17          | [7.73, 8.61]  |
| ATR-SLR       | 0.421      | [0.383, 0.457] | ATR-ALEPH     | 7.85          | [7.16, 8.55]  |
|               | mWER       | mPER       | GTM           | METEOR        |               |               |
| NLPR          | 0.416      | 0.326      | CMU           | 0.642         | NLPR          | 0.721         |
| CMU           | 0.420      | 0.337      | NLPR          | 0.626         | CMU           | 0.706         |
| ATR-C3        | 0.439      | 0.373      | ATR-C3        | 0.590         | ATR-C3        | 0.685         |
| ATR-ALEPH     | 0.454      | 0.418      | ATR-ALEPH     | 0.553         | ATR-SLR       | 0.642         |
| ATR-SLR       | 0.518      | 0.422      | ATR-SLR       | 0.547         | ATR-ALEPH     | 0.634         |

### Mixed Case Evaluation

|               | BLEU Score | NIST Score |               |               |               |               |
|---------------|------------|------------|---------------|---------------|---------------|---------------|
| ATR-ALEPH     | 0.478      | [0.444, 0.513] | ATR-ALEPH     | 7.65          | [7.03, 8.19]  |
| NLPR          | 0.409      | [0.383, 0.437] | NLPR          | 7.57          | [7.17, 7.95]  |
|               | mWER       | mPER       | GTM           | METEOR        |               |               |
| ATR-ALEPH     | 0.408      | 0.376      | ATR-ALEPH     | 0.600         | NLPR          | 0.721         |
| NLPR          | 0.546      | 0.482      | NLPR          | 0.492         | ATR-ALEPH     | 0.634         |
### B.3 Translation of ASR output Chinese to English – Automatic Evaluation

#### Supplied Data

|                  | Standard Evaluation | NIST Score |
|------------------|---------------------|------------|
| **RWTH**         | 0.383               | [7.19, 7.91] |
| **CMU**          | 0.363               | (6.94, 7.81) |
| **MIT-LL/AFRL**  | 0.360               | (6.68, 7.46) |
| **IBM**          | 0.336               | (6.01, 7.04) |

|                  | mWER | mPER | GTM | METEOR |
|------------------|------|------|-----|--------|
| **MIT-LL/AFRL**  | 0.560| 0.455| MIT-LL/AFRL | 0.525 |
| **RWTH**         | 0.565| 0.472| RWTH  | 0.488 |
| **CMU**          | 0.581| 0.499| CMU   | 0.483 |
| **IBM**          | 0.598| 0.504| IBM   | 0.481 |

(No mixed case submission)

#### Supplied Data + Tools

|                  | Standard Evaluation | NIST Score |
|------------------|---------------------|------------|
| **IBM**          | 0.358               | [5.81, 6.58] |
| **ATR-SLR**      | 0.267               | (5.22, 6.33) |

|                  | mWER | mPER | GTM | METEOR |
|------------------|------|------|-----|--------|
| **IBM**          | 0.596| 0.524| IBM | 0.471 |
| **ATR-SLR**      | 0.645| 0.547| ATR-SLR | 0.421 |

(No mixed case submission)

#### Unrestricted Data

|                  | Standard Evaluation | NIST Score |
|------------------|---------------------|------------|
| **IBM**          | 0.370               | [5.64, 6.42] |

|                  | mWER | mPER | GTM | METEOR |
|------------------|------|------|-----|--------|
| **IBM**          | 0.585| 0.519| 0.477| 0.495 |

(No mixed case submission)

#### C-STAR Data

|                  | Standard Evaluation | NIST Score |
|------------------|---------------------|------------|
| **NLPR**         | 0.385               | [7.65, 8.39] |
| **ATR-SLR**      | 0.340               | (6.27, 7.18) |

|                  | mWER | mPER | GTM | METEOR |
|------------------|------|------|-----|--------|
| **NLPR**         | 0.579| 0.477| NLPR| 0.507 |
| **ATR-SLR**      | 0.620| 0.528| ATR-SLR | 0.462 |

#### Mixed Case Evaluation

|                  | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|------------------|------------|------------|------|------|-----|--------|
| **NLPR**         | 0.298      | [5.64, 6.42] | 0.651| 0.574| 0.406| 0.580 |
### B.4 Translation of manual transcription Japanese to English – Automatic Evaluation

#### Supplied Data

|                | BLEU Score | NIST Score    | mWER | mPER | GTM | METEOR |
|----------------|------------|---------------|------|------|-----|--------|
| ITC-IRST       | 0.431      | [0.391, 0.471]| CMU  | 8.00 |     |        |
| RWTH           | 0.408      | [0.370, 0.443]| NTT  | 7.97 |     |        |
| CMU            | 0.393      | [0.361, 0.425]| RWTH | 7.86 |     |        |
| EDINBURGH      | 0.378      | [0.340, 0.414]| ATR-C3 | 7.74 |     |        |
| ATR-C3         | 0.374      | [0.338, 0.412]| ITC-IRST | 7.10 |     |        |
| NTT            | 0.345      | [0.314, 0.376]| USC-ISI | 4.87 |     |        |
| USC-ISI        | 0.283      | [0.251, 0.314]| EDINBURGH | 4.08 |     |        |

|                | BLEU Score | NIST Score    | mWER | mPER | GTM | METEOR |
|----------------|------------|---------------|------|------|-----|--------|
| ITC-IRST       | 0.516      | ITC-IRST     | 0.435| ITC-IRST | 0.492| NTT  | 0.603 |
| RWTH           | 0.536      | RWTH         | 0.444| RWTH   | 0.486| ATR-C3 | 0.601 |
| CMU            | 0.547      | CMU          | 0.455| ATR-C3 | 0.482| ITC-IRST | 0.587 |
| EDINBURGH      | 0.549      | ATR-C3      | 0.457| EDINBURGH | 0.475| RWTH | 0.586 |
| ATR-C3         | 0.557      | NTT          | 0.480| NTT    | 0.475| CMU  | 0.584 |
| NTT            | 0.595      | EDINBURGH   | 0.486| CMU    | 0.474| EDINBURGH | 0.517 |
| USC-ISI        | 0.623      | USC-ISI     | 0.521| USC-ISI | 0.448| USC-ISI | 0.494 |

#### Mixed Case Evaluation

|                | BLEU Score | NIST Score    | mWER | mPER | GTM | METEOR |
|----------------|------------|---------------|------|------|-----|--------|
| ITC-IRST       | 0.238      | [0.392, 0.461]| 6.99 | [6.58, 7.46]| 0.477| 0.409 | 0.533 | 0.567 |

#### Supplied Data + Tools

|                | BLEU Score | NIST Score    | mWER | mPER | GTM | METEOR |
|----------------|------------|---------------|------|------|-----|--------|
| ATR-C3         | 0.477      | [0.438, 0.516]| ATR-C3 | 8.17 |     |        |
| MICROSOFT      | 0.406      | [0.376, 0.439]| MICROSOFT | 8.04 |     |        |
| ATR-SLR        | 0.388      | [0.348, 0.431]| UTOKYO | 7.85 |     |        |
| UTOKYO         | 0.372      | [0.341, 0.402]| NGKUT  | 7.72 |     |        |
| NGKUT          | 0.342      | [0.312, 0.373]| ATR-SLR | 4.39 |     |        |
| USC-ISI        | 0.274      | [0.238, 0.309]| USC-ISI | 2.96 |     |        |

|                | BLEU Score | NIST Score    | mWER | mPER | GTM | METEOR |
|----------------|------------|---------------|------|------|-----|--------|
| ATR-C3         | 0.435      | ATR-C3       | 0.374| MICROSOFT | 0.583| ATR-C3 | 0.666 |
| MICROSOFT      | 0.516      | MICROSOFT    | 0.431| ATR-C3 | 0.552| UTOKYO | 0.621 |
| UTOKYO         | 0.531      | UTOKYO       | 0.440| MICROSOFT | 0.494| MICROSOFT | 0.620 |
| ATR-SLR        | 0.563      | NGKUT        | 0.467| NGKUT  | 0.470| NGKUT  | 0.603 |
| NGKUT          | 0.590      | ATR-SLR      | 0.519| ATR-SLR | 0.432| ATR-SLR | 0.521 |
| USC-ISI        | 0.665      | USC-ISI      | 0.573| USC-ISI | 0.406| USC-ISI | 0.429 |

#### Mixed Case Evaluation

|                | BLEU Score | NIST Score    | mWER | mPER | GTM | METEOR |
|----------------|------------|---------------|------|------|-----|--------|
| MICROSOFT      | 0.347      | [0.322, 0.372]| 7.01 | [6.66, 7.35]| 0.523| 0.452 | 0.508 | 0.620 |
### Unrestricted Data

#### Standard Evaluation

|            | BLEU Score | NIST Score       |
|------------|------------|------------------|
| NTT        | 0.393      | [0.359, 0.426]   |
| OKI        | 0.264      | [0.236, 0.294]   |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| NTT  | 0.559| NTT | 0.443  |
| OKI  | 0.607| OKI | 0.506  |

#### Mixed Case Evaluation

|            | BLEU Score | NIST Score       |
|------------|------------|------------------|
| OKI        | 0.289      | [0.262, 0.317]   |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| OKI  | 0.543| 0.466| 0.464  |
|      |      |      | 0.545  |

### C-STAR Data

#### Standard Evaluation

|            | BLEU Score | NIST Score       |
|------------|------------|------------------|
| RWTH       | 0.776      | [0.741, 0.809]   |
| ATR-SLR    | 0.727      | [0.689, 0.762]   |
| ATR-C3     | 0.687      | [0.648, 0.731]   |
| ATR-ALEPH  | 0.593      | [0.554, 0.635]   |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| RWTH | 0.243| RWTH| 0.186  |
| ATR-C3| 0.277| ATR-C3| 0.229 |
| ATR-SLR| 0.289| ATR-SLR| 0.244 |
| ATR-ALEPH| 0.361| ATR-ALEPH| 0.323 |

#### Mixed Case Evaluation

|            | BLEU Score | NIST Score       |
|------------|------------|------------------|
| ATR-ALEPH  | 0.592      | [0.554, 0.629]   |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
|      |      | RWTH| 0.787  |
|      |      | ATR-C3| 0.716 |
|      |      | ATR-SLR| 0.693 |
|      |      | ATR-ALEPH| 0.607 |
|      |      |      | 0.854  |
|      |      |      | 0.810  |
|      |      |      | 0.800  |
|      |      |      | 0.720  |

### GTM and METEOR

- **Standard Evaluation**
  - GTM: [0.236, 0.294]
  - METEOR: [7.03, 7.70]

- **Mixed Case Evaluation**
  - GTM: [0.262, 0.317]
  - METEOR: [7.07, 7.35]
### B.5 Translation of ASR output Japanese to English – Automatic Evaluation

#### Supplied Data

|                  | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|------------------|------------|------------|------|------|-----|--------|
| ITC-IRST         | 0.430      | [0.393, 0.468] | RWTH | 8.53 |     | [8.06, 8.96] |
| RWTH             | 0.427      | [0.392, 0.460] | NTT  | 8.32 | [7.93, 8.67] |
| NTT              | 0.375      | [0.340, 0.405] | ITC-IRST | 8.27 | [7.82, 8.71] |

|                  | mWER | mPER | GTM | METEOR |
|------------------|------|------|-----|--------|
| ITC-IRST         | 0.507 | RWTH | 0.412 | ITC-IRST | 0.504 | NTT | 0.633 |
| RWTH             | 0.512 | ITC-IRST | 0.419 | RWTH | 0.496 | RWTH | 0.620 |
| NTT              | 0.564 | NTT | 0.457 | NTT | 0.487 | ITC-IRST | 0.618 |

#### Mixed Case Evaluation

|                  | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|------------------|------------|------------|------|------|-----|--------|
| ITC-IRST         | 0.435      | [0.399, 0.470] | 7.87 | 7.51 | 8.27 | 0.574 | 0.537 | 0.618 |

#### Supplied Data + Tools

|                  | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|------------------|------------|------------|------|------|-----|--------|
| ATR-SLR          | 0.383      | [0.342, 0.421] | UTOKYO | 7.42 | [6.91, 7.87] |
| UTOKYO           | 0.336      | [0.305, 0.370] | ATR-SLR | 4.27 | [3.68, 4.87] |

|                  | mWER | mPER | GTM | METEOR |
|------------------|------|------|-----|--------|
| UTOKYO           | 0.568 | UTOKYO | 0.472 | UTOKYO | 0.469 | UTOKYO | 0.597 |
| ATR-SLR          | 0.574 | ATR-SLR | 0.531 | ATR-SLR | 0.423 | ATR-SLR | 0.513 |

(No mixed case submission)

#### Unrestricted Data

(No submissions)

#### C-STAR Data

|                  | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|------------------|------------|------------|------|------|-----|--------|
| ATR-SLR          | 0.679      | [0.643, 0.715] | 10.04 | 9.40 | 10.65 | 0.324 | 0.282 | 0.671 | 0.761 |

(No mixed case submission)
## B.6 Translation of manual transcription Arabic to English – Automatic Evaluation

### Supplied Data

| BLEU Score | NIST Score |
|-------------|------------|
| 0.573       | 9.78       |
| [0.537, 0.608] | [9.27, 10.27] |
| 0.562       | 9.66       |
| [0.525, 0.599] | [9.16, 10.12] |
| 0.547       | 9.33       |
| [0.517, 0.581] | [8.82, 9.84] |
| 0.538       | 9.27       |
| [0.503, 0.574] | [8.87, 9.63] |
| 0.533       | 8.74       |
| [0.495, 0.571] | [8.35, 9.11] |
| 0.511       | 8.62       |
| [0.476, 0.545] | [8.05, 9.19] |
| 0.446       | 8.14       |
| [0.411, 0.479] | [7.08, 8.22] |
| 0.409       | 6.54       |
| [0.382, 0.439] | [6.81, 7.30] |
| 0.374       | 2.65       |
| [0.335, 0.415] | [2.35, 3.41] |

### Standard Evaluation

| BLEU Score | NIST Score |
|-------------|------------|
| 0.350       | 0.683      |
| TALP-phrase | TALP-phrase |
| 0.303       | 0.303      |
| TALP-phrase | TALP-phrase |
| 0.669       | 0.733      |
| ITC-IRST    | ITC-IRST   |
| 0.313       | 0.313      |
| TALP-phrase | TALP-phrase |
| 0.656       | 0.708      |
| RWTH        | RWTH       |
| 0.319       | 0.319      |
| TALP-ngram  | TALP-ngram |
| 0.368       | 0.368      |
| TALP-phrase | TALP-phrase |
| 0.647       | 0.689      |
| ITC-IRST    | ITC-IRST   |
| 0.376       | 0.376      |
| CMU         | CMU        |
| 0.430       | 0.430      |
| USC-ISI     | USC-ISI    |
| 0.515       | 0.515      |
| USC-ISI     | USC-ISI    |
| 0.483       | 0.483      |
| USC-ISI     | USC-ISI    |
| 0.551       | 0.551      |
| USC-ISI     | USC-ISI    |
| 0.546       | 0.546      |

### Mixed Case Evaluation

| BLEU Score | NIST Score |
|-------------|------------|
| 0.576       | 9.38       |
| [0.546, 0.608] | [9.01, 9.75] |
| 0.545       | 8.52       |
| [0.515, 0.579] | [8.07, 8.95] |
| 0.320       | 0.702      |
| ITC-IRST    | ITC-IRST   |
| 0.277       | 0.732      |
| IBM         | IBM        |
| 0.334       | 0.680      |
| IBM         | IBM        |
| 0.303       | 0.689      |

### Supplied Data + Tools

| BLEU Score | NIST Score |
|-------------|------------|
| 0.560       | 9.59       |
| [0.525, 0.598] | [9.10, 10.04] |
| 0.396       | 5.05       |
| [0.357, 0.431] | [4.05, 5.94] |
| 0.357       | 0.666      |
| IBM         | IBM        |
| 0.309       | IBM        |
| 0.666       | 0.712      |
| USC-ISI     | USC-ISI    |
| 0.469       | USC-ISI    |
| 0.560       | USC-ISI    |
| 0.562       | USC-ISI    |

### Standard Evaluation

| BLEU Score | NIST Score |
|-------------|------------|
| 0.571       | 9.21       |
| [0.571, 0.604] | [8.83, 9.58] |
| 0.318       | 0.280      |
| 0.695       | 0.712      |

### Mixed Case Evaluation
### Unrestricted Data

|               | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|---------------|------------|------------|------|------|-----|--------|
| IBM           | 0.600      | [0.565, 0.633] | IBM  | 9.76 | [9.28, 10.22] |
| NTT           | 0.472      | [0.438, 0.506] | NTT  | 9.38 | [8.94, 9.82]  |
| mWER          |            |            |      |      |     |        |
| IBM           | 0.333      | IBM        | 0.294 | IBM  | 0.682 | IB      |
| NTT           | 0.484      | NTT        | 0.377 | NTT  | 0.621 | NTT    |

### Mixed Case Evaluation

|               | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|---------------|------------|------------|------|------|-----|--------|
| IBM           | 0.604      | [0.575, 0.635] | 9.37 | [8.97, 9.75] | 0.295 | 0.264 | 0.712 | 0.726 |

### C-STAR Data

|               | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|---------------|------------|------------|------|------|-----|--------|
| ATR-ALEPH     | 0.382      | [0.348, 0.417] | 6.22 | [5.62, 6.83] | 0.527 | 0.498 | 0.481 | 0.543 |

### Mixed Case Evaluation

|               | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|---------------|------------|------------|------|------|-----|--------|
| ATR-ALEPH     | 0.382      | [0.348, 0.417] | 6.23 | [5.72, 6.74] | 0.471 | 0.446 | 0.542 | 0.543 |
## B.7 Translation of manual transcription Korean to English – Automatic Evaluation

### Supplied Data

|                  | Standard Evaluation | NIST Score          |
|------------------|---------------------|---------------------|
| **BLEU Score**   |                     | [0.330, 0.405]     |
| EDINBURGH        | 0.367               | CMU                 |
| CMU              | 0.358               | NTT                 |
| NTT              | 0.307               | USC-ISI             |
| USC-ISI          | 0.237               | EDINBURGH           |

|                  | mWER    | mPER    | GTM     | METEOR  |
|------------------|---------|---------|---------|---------|
| EDINBURGH        | 0.557   | 0.444   | CMU     |
| CMU              | 0.561   | EDINBURGH| NTT    |
| NTT              | 0.645   | NTT     | EDINBURGH|
| USC-ISI          | 0.678   | USC-ISI | NTT     |

(No mixed case submission)

### Supplied Data + Tools

|                  | Standard Evaluation | NIST Score          |
|------------------|---------------------|---------------------|
| **BLEU Score**   |                     | [0.221, 0.282]     |
| USC-ISI          | 0.252               | SEHDA               |
| SEHDA            | 0.206               | USC-ISI             |

|                  | mWER    | mPER    | GTM     | METEOR  |
|------------------|---------|---------|---------|---------|
| USC-ISI          | 0.659   | USC-ISI | SEHDA  |
| SEHDA            | 0.703   | SEHDA   | SEHDA  |

(No mixed case submission)

### Unrestricted Data

|                  | Standard Evaluation | NIST Score          |
|------------------|---------------------|---------------------|
| **BLEU Score**   |                     | [0.316, 0.381]     |
| NTT              | 0.350               |                    |

(No mixed case submission)

### C-STAR Data

|                  | Standard Evaluation | NIST Score          |
|------------------|---------------------|---------------------|
| **BLEU Score**   |                     | [0.374, 0.449]     |
| ATR-ALEPH        | 0.412               |                    |

|                  | mWER    | mPER    | GTM     | METEOR  |
|------------------|---------|---------|---------|---------|
| ATR-ALEPH        | 0.411   | 0.447   | 0.499   | 0.563   |

### Mixed Case Evaluation

|                  | Standard Evaluation | NIST Score          |
|------------------|---------------------|---------------------|
| **BLEU Score**   |                     | [0.377, 0.445]     |
| ATR-ALEPH        | 0.411               |                    |

|                  | mWER    | mPER    | GTM     | METEOR  |
|------------------|---------|---------|---------|---------|
| ATR-ALEPH        | 0.411   | 0.447   | 0.499   | 0.563   |
## B.8 Translation of manual transcription English to Chinese – Automatic Evaluation

### Supplied Data

| Evaluation 1 - ASR segmentation based | BLEU Score | NIST Score |
|--------------------------------------|------------|------------|
| EDINBURGH                            | 0.213      | [4.91, 5.44] |
| RWTH                                 | 0.200      | [4.84, 5.35] |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| RWTH | 0.612 | EDINBURGH | 0.527 |
| EDINBURGH | 0.620 | RWTH | 0.529 |

| Evaluation 2 - Character segmented | BLEU Score | NIST Score |
|------------------------------------|------------|------------|
| EDINBURGH                          | 0.301      | [5.80, 6.44] |
| RWTH                               | 0.288      | [5.70, 6.35] |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| EDINBURGH | 0.558 | EDINBURGH | 0.445 |
| RWTH | 0.560 | RWTH | 0.446 |

### Supplied Data + Tools

| Evaluation 1 - ASR segmentation based | BLEU Score | NIST Score |
|--------------------------------------|------------|------------|
| MICROSOFT                            | 0.206      | [4.97, 5.50] |
| RWTH                                 | 0.191      | [4.71, 5.22] |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| MICROSOFT | 0.613 | MICROSOFT | 0.520 |
| RWTH | 0.633 | RWTH | 0.546 |

| Evaluation 2 - Character segmented | BLEU Score | NIST Score |
|------------------------------------|------------|------------|
| MICROSOFT                          | 0.306      | [6.14, 6.66] |
| RWTH                               | 0.282      | [5.68, 6.27] |

| mWER | mPER | GTM | METEOR |
|------|------|-----|--------|
| MICROSOFT | 0.548 | MICROSOFT | 0.430 |
| RWTH | 0.564 | RWTH | 0.450 |

### Unrestricted Data

(No submissions)

### C-STAR Data

| Evaluation 1 - ASR segmentation based | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|--------------------------------------|------------|------------|------|------|-----|--------|
| ATR-ALEPH                            | 0.098 [0.078, 0.118] | 3.03 [2.74, 3.31] | 0.798 | 0.746 | 0.363 |

| Evaluation 2 - Character segmented | BLEU Score | NIST Score | mWER | mPER | GTM | METEOR |
|------------------------------------|------------|------------|------|------|-----|--------|
| ATR-ALEPH                          | 0.183 [0.161, 0.207] | 4.08 [3.68, 4.45] | 0.725 | 0.646 | 0.450 |