Towards Innovative Evaluation Methodologies for Speech Translation

Michael PAUL Hiromi NAKAIWA
ATR Spoken Language Translation Labs
Keihanna Science City
619-0288 Kyoto, Japan
Michael.Paul@atr.jp Hiromi.Nakaiwa@atr.jp

Marcello FEDERICO
ITC-irst - Centro per la Ricerca
Scientifica e Technologica
I-38050 Povo-Trento, Italy
federico@itc.it

Abstract

This paper reports on activities within the C-STAR[1] consortium which aim at novel speech translation technologies and their evaluation. In C-STAR, current state-of-the-art speech translation systems developed by the partners are evaluated and discussed on a regular basis by means of evaluation workshops. The objectives of these workshops are to provide a framework for the validation of existing evaluation methodologies concerning their applicability to the evaluation of speech translation technologies, and to open new directions on how to improve current methods.
Keywords: *C-STAR, evaluation methodologies, IWSLT, multilingual corpus, speech translation*

1 Introduction

Speech translation technologies attempt to cross the language barriers between people having different native languages who want to engage in conversation by using their mother-tongue. The importance of these technologies is increasing because there are many more opportunities for cross-language communication in face-to-face and telephone conversation. Other applications of speech translation technologies include cross-language information retrieval systems that allow users to access information in a foreign language by using their native language.

Novel technologies have been proposed to tackle the problems in spoken language translation research. A number of institutes are developing huge bilingual or multilingual speech corpora. Machine translation (MT) technologies based on machine learning, such as statistical MT and example-based MT, are being applied to the translation of spoken language by using these corpora. However, there is still no concrete standard methodology for comparing the translation quality of speech translation systems.

One of the prominent research activities in spoken language translation is the work being conducted by the C-STAR consortium, which is an international partnership of research laboratories engaged in the automatic translation of spoken language. Current members include ATR (Advanced Telecommunications Research Institute, Japan), CAS (Chinese Academy of Sciences, China), CLIPS (University Joseph Fourier, France), CMU (Carnegie Mellon University, USA), ETRI (Electronics and Telecommunications Research Institute, Korea), ITC-irst (Center for Scientific and Technological Research, Italy), and UKA (University of Karlsruhe, Germany). One of C-STAR's ongoing projects is the joint development of a speech corpus that handles a common task in multiple languages. As a first result of this activity, a Japanese-English speech corpus comprising tourism-related sentences, originally compiled by ATR, has been translated into the native languages of the C-STAR members.

The corpus, described in detail in Section 2, serves as a primary source for developing and evaluating broad-coverage speech translation technologies. They will be evaluated and discussed on a regular basis by means of evaluation workshops (cf. Section 3). The next workshop takes place in 2004 and is open to external participants. The corpus supplied for this year's conference, the reference translations, the output of the participating MT systems, and the evaluation results will be made publicly available after the workshop (cf. Section 4). These resources can be used as a benchmark for future research on MT systems and MT evaluation methodologies.

2 Multilingual Spoken Language Corpus

The multilingual spoken language corpus, jointly developed by the C-STAR partners, is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country and cover utterances for every potential subject in travel situations for several European and Asiatic languages. The initial collection of Japanese and English sentence pairs is being translated into Chinese, Korean, and (partially) Italian, and will be extended further to Spanish, French, and German. The statistics of

[1] Consortium for Speech Translation Advanced Research, http://www.c-star.org/

the *Basic Travel Expressions Corpus* (BTEC*)[2] shared between C-STAR partners are summarized in Table 1.

Table 1. BTEC* corpus

language	sentence count	word tokens	word types	words per sentence
Japanese	162K	1,114,186	18,781	6.9
English		952,300	12,404	5.9
Chinese		959,846	15,516	5.9
Korean		1,211,129	21,837	7.5
Italian	48K	361,250	14,871	7.4

Word token refers to the number of words in the corpus, whereas *word type* refers to the vocabulary size. Table 2 gives some examples of the English BTEC*.

Table 2. English sample sentences

I want to buy a roll of fi lm.
I'd like to reserve a table for eight.
Do you have some tea?
I'd like to return the car.
You need to cross the bridge to go there.
My friend was hit by a car and badly injured.
I do not like the color.

In C-STAR, the BTEC* corpus serves as a primary source for developing and evaluating broad-coverage speech translation technologies. Due to its multilinguality, BTEC* has great potential for the development of MT technologies from many-to-many languages. However, BTEC* sentences are not transcriptions of actual interactions, but were generated by experts to cover utterances for potential subjects in travel situations. Thus the corpus may have the following problems: (1) BTEC* may lack utterances that appear in real interactions; (2) the frequency distribution of BTEC* may be different from the "actual" one.

In order to get an idea of how "realistic" the BTEC* corpus is, we compared the Japanese and English parts of this corpus with a dialogue corpus containing simulated (role play) dialogues (MAD) between two native speakers of different languages [4]. The size of the BTEC* corpus is around twenty times the size of the MAD corpus (514 dialogues, 6,972 utterances, 54,452 word tokens, 4,192 word types). Its average number of words per sentence (J: 6.9, E: 5.9) is much shorter than that for MAD (J: 10.0, E: 10.3). The reason for this is that simulated dialogues contain more complex or compound sentences (BTEC*: 17.2%, MAD: 31.7%) as well as many modifiers, e.g., adverbs and adjectives, to refer to actual situations. However, an examination of how much BTEC* covers MAD revealed that BTEC* is still useful for the development of speech translation technologies. Figure 1 shows the BTEC* coverage of MAD N-grams.

[2]Up-to-date information about the BTEC* corpus can be found at http://cstar.atr.jp/cstar-corpus

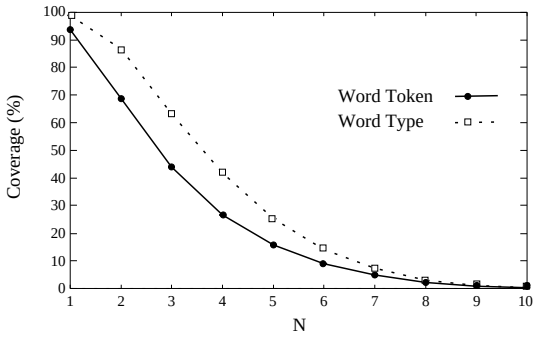

Figure 1. N-gram coverage (Japanese)

At a glance, BTEC* tri-grams cover 63.1% of the MAD tri-gram tokens. Although we should consider the difference in corpus size, we can conclude that BTEC* covers local word sequences in the MAD corpus well. In addition, we counted how many sentences of MAD are covered by BTEC* focusing on either content words or functional words. The results (content words: 21.0% of word tokens, 18.6% of word types; functional words: 46.9% of word tokens, 44.1% of word types) reveal that there is still a need for the collection of more actual utterances in order to create a quantitatively correct sample of reality.

Therefore, we are investigating paraphrasing existing corpora in order to achieve a broader coverage and recording actual interactions to improve the development of corpus-based speech translation technologies for real use [4].

3 Evaluation Methodologies

Traditionally, it is difficult to quantify what good translations are. Despite extensive research efforts, there are no universally accepted and reliable evaluation metrics for the evaluation of machine translation output. Early studies on the evaluation of machine output assessed both the *quality* and *informativeness* of the output [1], [6]. The quality of machine output, i.e., the understandability of a machine translation by a native speaker of the target language, can be judged using subjective gradings that characterize various output disfluencies. The informativeness criteria judges the translation output against the input, i.e., it evaluates whether it preserves information in the source text, while adding no new information. However, such an approach requires bilingual evaluators and leads to an increase in evaluation costs. Another way of measuring informativeness is to compare against other outputs. If those reference translations are produced by humans, an informativeness measure can assess the accuracy of coverage of information in the reference output. Therefore, recent competitive MT evaluations, like the series of DARPA MT evaluations in the mid 1990's [10], evaluate machine translation output with human reference translations on the basis of *fluency* and *adequacy* [5]. *Fluency* refers to the degree to

which the translation is well-formed according to the grammar of the target language. *Adequacy* refers to the degree to which the translation communicates the information present in the reference output. The fluency and adequacy judgments consist of one of the grades listed in Table 3.

Table 3. Human assessment

Fluency	
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

Adequacy	
5	All Information
4	Most Information
3	Much Information
2	Little Information
1	None

A problem with evaluation methods using reference translations is its potential incompleteness, i.e., there is usually more than one correct translation of a specific input sentence. Moreover, human judgments are quite expensive and time consuming.

Therefore, recent research efforts, like the TIDES program[3], focus on automatic evaluation using multiple reference translations, whereby subjective evaluations are intended to ground automatic evaluation measures in human judgments.

The increasing availability of bilingual resources and new ideas in data-driven MT research led to various automatic scoring metrics that allow us to compare different MT systems and to monitor the progress of MT system development.

- BLEU: the geometric mean of n-gram precision by the system output with respect to reference translations [9].
- NIST: a variant of BLEU using the arithmetic mean of weighted n-gram precision values [2].
- *Multiple Word Error Rate* (mWER): the edit distance between the system output and the closest reference translation [7].
- *Position independent mWER* (mPER): a variant of mWER which disregards word ordering [8].

Excluding NIST, the scores of all automatic evaluation metrics are in the range of [0,1]. NIST is always positive and its scoring range does not have a theoretical upper limit. In contrast to mWER and mPER, higher BLEU and NIST scores indicate better translations.

4 C-STAR Evaluation Campaigns

Within the C-STAR consortium the decision was taken to organize, on a regular basis, speech translation evaluation campaigns and workshops, mainly focusing on speech translation research and evaluation. Activities within C-STAR include the development of a large multilingual parallel corpus, as described in Section 2, to be used for common evaluations.

[3]http://www.ldc.upenn.edu/Projects/TIDES/tidesmt.html

4.1 Evaluation Campaign 2003

The first internal workshop utilizing the BTEC* corpus took place in 2003 and was restricted to C-STAR members only. It concentrated on assessing text translation algorithms on the domain of tourism [3].

The translation directions were from the native languages of the C-STAR partners (Chinese, Italian, Japanese, and Korean) into English. The training data consisted of a fixed number of English sentences provided with translations into the respective source language. Participants were allowed to use any additional monolingual resources, e.g. text corpora, grammars, word lists, segmentation tools. The test data consisted of 500 sentences from the BTEC* corpus reserved for evaluation purposes.

We used the evaluation metrics described in Section 3 to evaluate six MT systems (MT_1, ... ,MT_6) developed by the partners, whereby up to 16 multiple references were used for the calculation of the automatic evaluation scores. The evaluation results[4] are summarized in Table 4.

Table 4. Evaluation results

metric	MT_1	MT_2	MT_3	MT_4	MT_5	MT_6
fluency	3.76	4.04	2.81	2.30	3.74	-
adequacy	4.00	3.92	3.01	2.59	3.22	-
BLEU	0.631	0.643	0.280	0.284	0.610	0.410
NIST	11.191	9.982	6.604	6.197	3.171	8.914
mWER	0.272	0.283	0.573	0.561	0.456	0.466
mPER	0.224	0.254	0.476	0.471	0.445	0.371

Due to different source languages[5], a direct comparison between all system outputs might be problematic. However, the utilized evaluation measurements rely only on target language information (MT output and target references), thus the scores should be directly comparable even if the corresponding MT tasks are not comparable.

The obtained results reveal some inconsistencies of automatic scoring methods concerning the ranking of MT systems. If we take the harmonic mean of the fluency and adequacy judgments, MT_2 seems to perform slightly better than MT_1. However, excluding BLEU, all automatic evaluation metrics prefer MT_1. Moreover, systems with lower performance (MT_3, MT_4, and MT_5) are also misranked by all metrics. This indicates that the utilized evaluation schemes cannot distinguish well between systems of similar performance and thus opposes the findings of previous studies [2]. Another scoring anomaly can be found for system MT_5, which gets the worst NIST score, but whose BLEU score is the highest one out of the three Chinese-to-English MT systems. Both schemes penalize translations which are shorter than the reference

[4]MT_6 was evaluated only automatically.

[5]Only three systems (MT_4, MT_5, and MT_6) used the same language, i.e., Chinese, as its input.

translations using a multiplicative *brevity penalty* factor. BLEU penalizes more than NIST when translations of the MT system are slightly shorter than the reference translations. However, the shorter the system translations, the more sensibly NIST penalizes compared to BLEU. In the case of MT_5, we observed that, on average, the translations are significantly shorter than the reference translations. Table 5 illustrates the effect of different brevity penalty factors on the automatic evaluation scores. In the case of BLEU, the system score is reduced to half, whereas the NIST score is punished much more harshly, i.e., it is reduced to less than 1/3.

Table 5. Brevity penalty factor (BP)

system	sys/ref length ratio	BP of BLEU	BP of NIST
MT_4	0.98	0.98	1.00
MT_5	0.58	0.48	0.29
MT_6	1.13	1.00	1.00

However, depending on how we compare MT outputs with reference translations, we obtain quite different results, because each MT system has its own *style* of outputting translations. These changes in automatic evaluation scores are caused mainly by the following factors: (1) *case-sensitiveness* (lower-case only vs. mixed); (2) *punctuation marks* (with or without); (3) *writing style* alternations, like numerals (spelled-out vs. digits), time/date expressions ("eleven thirty" vs. "half past eleven", "july eigth" vs. "eighth of july"), abbreviations ("o.k." vs. "okay" vs. "OK"), and word compounds ("duty-free" vs. "duty free"); (4) *level of granularity*, i.e., comparison using words only, words with part-of-speech tags, words and their inflectional attributes. The evaluation parameters used for the evaluation results given in Table 4 are (1) case-insensitive; (2) punctuation marks are ignored; (3) comparison of surface words only; (4) numerals are spelled-out.

In order to verify the dependency of automatic evaluation scores from evaluation parameters, we applied different evaluation parameter settings. Table 6 illustrates to what degree evaluation parameters do influence the automatic scoring of different MT systems. The numbers show the dynamics, i.e., the amplitude of variation, of each score, obtained by applying several evaluation parameter settings on the output of a given system.

Table 6. Automatic scoring variation

metric	MT_1	MT_2	MT_3	MT_4	MT_5	MT_6
BLEU	0.066	0.134	0.059	0.131	0.070	0.061
NIST	1.124	3.914	0.585	1.759	0.034	1.035
mWER	0.327	0.226	0.069	0.263	0.077	0.249
mPER	0.326	0.237	0.126	0.274	0.084	0.270

The largest variations concerning BLEU and NIST scores can be seen for MT_2, whose translations were case-insensitive and without punctuation marks. Therefore, the comparison with cased reference translations drastically affects the automatic evaluation scores. On the other hand, the imprecise generation of punctuation marks by system MT_1 results in much higher word error rates when this evaluation parameter is used.

Such drastic differences in the scoring results lead to discrepancies in MT system rankings when multiple MT systems are to be compared. However, the judgment of the MT output quality should be independent of system specific features. Moreover, the selection of the evaluation parameter depends on the evaluation task. Whereas orthographic features (case, punctuation marks, etc.) are important for written text, they might be less relevant for the evaluation of spoken language. Therefore, we would like to investigate in more detail whether current evaluation metrics are suitable for the task of speech translation, and open new directions on how to improve current methods.

4.2 Evaluation Campaign 2004

In order to achieve these goals, this year's workshop[6] will be open to external participants and focus on the validation of existing evaluation methodologies concerning their applicability to the evaluation of speech translation technologies.

The Evaluation Campaign 2004 will be carried out using parts of the multilingual BTEC* corpus. This involves the translation of source language sentences (Chinese and Japanese) into the target language (English). Participants will be supplied with 20,000 sentence pairs for each translation direction (Chinese-to-English, Japanese-to-English). These training sentences are randomly selected from the BTEC* corpus. Word segmentations for the Chinese and Japanese subsets are provided, in case appropriate tools are not available for a participant. The test set consists of 500 sentences randomly selected from parts of the BTEC* corpus reserved for evaluation purposes that are different from those used in the previous evaluation campaign.

We distinguish three different language resource conditions. The training data of the *Small Data* track is limited to the supplied corpus only. The *Additional Data* track limits the use of bilingual resources to those that are publicly available from the LDC[7] (Chinese-to-English only). No restrictions on linguistic resources are imposed for the *Unrestricted Data* track. Separate run submissions are required for each track, whereby each participant can submit multiple runs for the same track. However, only the first submitted run will be evaluated by human judges.

The MT output has to be confirmed with the following evaluation parameters: (1) *case-insensitive*, i.e.,

[6] International Workshop on Spoken Language Translation, http://www.slt.atr.jp/IWSLT2004

[7] Linguistic Data Consortium, http://www.ldc.upenn.edu/

lower-case only; (2) *no punctuation marks*.

The translation quality will be measured using both human assessments and automatic scoring techniques. The subjective evaluation is carried out by native speakers of American English. The translation quality is judged based on the *fluency* and *adequacy* of the translation similar to the evaluation guidelines used in the TIDES program. In order to minimize grading inconsistencies between evaluators due to contextual misinterpretations of the translations, the situation in which the sentence is uttered (corpus annotations like "sightseeing" or "restaurant") will be provided for the adequacy judgment. Each translation of a single MT system will be evaluated by at least three judges. The automatic evaluation is carried out using the automatic scoring metrics introduced in Section 3, whereby we utilize up to 16 human reference translations.

The analysis of the evaluation results will be carried out by members of the Evaluation Committee, which also includes representatives of the participating organizations. The results will be published at the workshop to be held September 30 and October 1, 2004 in Kyoto, Japan. In addition, all language resources (supplied corpus, reference translations, MT output, evaluation results) will be made publicly available after the workshop.

5 Concluding Remarks

The analysis of the BTEC* corpus has shown that there is still a need for the collection of more actual utterances in order to create a quantitatively correct sample of reality. However, due to its multilinguality, BTEC* has great potential for the development of machine translation technologies from many-to-many languages.

A closed evaluation campaign based on the BTEC* corpus permitted us to compare MT across different source languages. The evaluation across different systems was possible only for Chinese-to-English. The results revealed some inconsistencies between MT system rankings depending on the utilized automatic scoring schemes as well as evaluation parameters. However, the outcomes of this evaluation should stimulate discussions about technical issues related to machine translation algorithms and how to improve automatic scoring methods. This year's workshop extends the evaluation framework to new language pairs and allows evaluation across a large number of different MT systems using the same training data. Most important, the obtained resources will be used as a benchmark for future research on MT systems and MT evaluation methodologies.

So far, we have focused on written text in utterance style and the applicability of current evaluation methodologies. Future evaluation workshops will focus on speech input to MT systems and evaluation factors inherent to speech-to-speech translation. In particular, we would like to investigate the robustness of MT systems for speech recognition errors, the processing time for real-time response, and the usability of speech-to-speech technologies (user-interface, end-to-end evaluation) in real situations.

Acknowledgment

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled, *"A study of speech dialogue translation technology based on a large corpus"*. The authors would like to thank all C-STAR members for their cooperation and helpful discussions.

References

[1] J. B. Carroll. An experiment in evaluating the quality of translations. In *Languages and machines: Computers in translation and linguistics — A Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council*, pages 67–75. National Research Council Publication 1416, Washington, DC, 1966. http://www.nap.edu/books/ARC000005/html/index.html.

[2] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the HLT 2002*, pages 257–258, San Diego, USA, 2002.

[3] M. Federico. Evaluation frameworks for speech translation technologies. In *Proc. of the EUROSPEECH03*, pages 377–380, Geneve, Switzerland, 2003.

[4] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In *Proc. of the EUROSPEECH03*, pages 381–384, Geneve, Switzerland, 2003.

[5] Linguistic Data Consortium. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Chinese-English Translations Revision 1.0*, 2002. http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf.

[6] M. Nagao, J. Tsujii, and J. Nakamura. The Japanese government project for machine translation. *Computational Linguistics*, 11(2-3):91–110, 1985.

[7] S. Niessen, F. J. Och, G. Leusch, and H. Ney. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proc. of the 2nd LREC*, pages 39–45, Athens, Greece, 2000.

[8] F. J. Och. Minimum error rate training in statistical machine translation. In *Proc. of the 41st ACL*, pages 160–167, Sapporo, Japan, 2003.

[9] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318, Philadelphia, USA, 2002.

[10] J. S. White, T. O'Connell, and F. O'Mara. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proc of the AMTA*, pages 193–205, 1994.