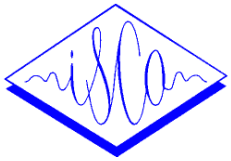

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

RELIABILITY MEASURES FOR TRANSLATION QUALITY

Eiichiro Sumita, Yasuhiro Akiba, and Kenji Imamura

ATR Spoken Language Research Laboratories
2-2 Hikaridai, Seika, Soraku
Kyoto 619-0288, JAPAN
eiichiro.sumita, yasuhiro.akiba, kenji.imamura@atr.co.jp

ABSTRACT

This paper proposes new reliability measures on translation quality. A reliability measure is important for a translation system to select the best one of multiple translations. The proposed measures are independent of translation systems, so they are applicable to any translation system. Experiments on selection proved the effectiveness of our measures.

1. INTRODUCTION

Most commercial machine translation systems are designed to generate as many translations as possible without taking the quality of each translation into consideration. Consequently, the output is a mixture of good translations and bad translations. There is no information on the reliability of each translation. This can be compared to a jar of cookies when some of them are poisoned and others are not: no one wants to eat out of it. In contrast, some recent research systems have been designed to output translation when the system is confident of the quality or can associate a reliability value to each translation. The user is safe by confining himself/herself to using translations with a high reliability value.

If a system consists of multiple translation engines, the system has to select the best one of multiple outputs. The system must have a reliability measure for translation selection. Therefore, reliability measures have been attracting great attention.

Section 2 reviews previous research on reliability measures, and Section 3 proposes two reliability measures. Section 4 reports on our experiment of translation selection, and Section 5 concludes the paper.

2. PPEVIOUS WORK

This section reviews the reliability measures proposed previously and explains the representative application, i.e., selecting the best one of translations produced by multi-engines.

2.1. Two types of reliability measures

There are two types of reliability measures: system-dependent and system-independent.

System-dependent measures are those that a system calculates by itself based on its knowledge and procedures. The similarity scores used in EBMT [8] and the probability used in SMT [6] are typical. They are well suited to each paradigm, but of course incomplete in that they are not reliable for every case. In addition, they have demerits in that they are only applicable to the concerned systems.

System-independent measures do not refer to system-internal information. The method using N-gram statistics of a target language corpus has been proposed. [4, 3] It is based on assumptions that (1) the fluency of translations are effective for selecting good translations because they are sensitive to the broken target sentences due to errors in translation processes, and (2) the source and target correspondences from the semantic point of view are often kept in a state-of-the-art translation system. However, the second assumption does not necessarily hold.

To overcome the demerits of N-gram and keep the merits of system-independent measures, we propose new measures that refer to both sides of translation.

2.2. Multi-engine translation

Multiple translations have been generated for several reasons, such as adoption of multiple engines [12, 18, 15], automated post-editing [5, 16], automated paraphrasing [7] and multiple interpretations in analysis.

As explained in a later section, we have built three different MT systems for the same domain [12], but their accuracies are different. Next we show a sample of different English translations obtained by the three systems for the same Japanese sentence J1. The brackets show the quality rank judged by a human translator according to the criteria of [13]: [A]: perfect; [B]: OK; [C]: understandable; [D]: BAD.

J1. o-shiharai wa genkin desu ka kurejitto kaado desu ka?
{payment/TOPIC/cash/be/QUESTION/credit/card/ be/QUESTION}
[B] Is the payment cash? Or is it the credit card?
[A] Would you like to pay in cash or with a credit card?
[C] Could you cash or credit card?

Furthermore, since each system has its own different well-translated sentences, the differences are substantial. Thus, we can obtain the large increase in accuracy by a “dream” MT as shown in Figure 1 if we can choose the best one of the three different translations for each input sentence.

This is often the case with multi-engines because different engines are dependent on different technologies, and the maturity of system elements are varied. Therefore, there is no guarantee that qualities of translations are in accord for every input. Therefore, the method to select the best is important for improving the performance of an integrated system.

Furthermore, selecting approach is an easy, quick and low-cost method of improving total performance because there is no need to investigate the messy relationships between resources and processes of the component systems.

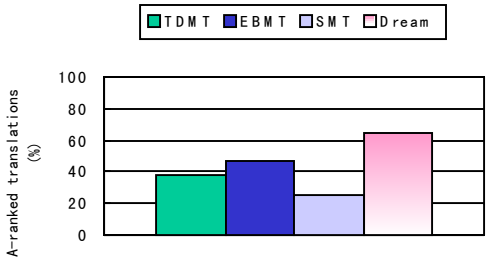

Figure 1 Component MTs vs. Dream MT

2.3. Selection framework

There are two usages of reliability measures: single-measure and multiple-measure.

Single-measure is a method to order multiple translations by using a single reliability measure. Several studies have revealed the effectiveness of this approach [3, 4].

Multiple-measure is important because we do not necessarily assume that a single reliability measure achieves the best performance in selecting translations, while we do assume that each reliability measure is sensitive to certain aspects of translation quality. Therefore, a combination of multiple measures is probably necessary to achieve higher accuracy. Two approaches have been proposed: (1) learning to rescale multiple measures [15]; (2) combining multiple measures with a decision tree learner [1, 18].

3. RELIABILITY MEASURES

Our measures are computed based on: (1) the system's input and output, i.e., pairs of the input sentence and its machine-generated translation, and (2) system external knowledge, i.e., a bilingual corpus.

We propose two measures in this paper (1) Sentence-Based Measure (SBM) and (2) Alignment-Based Measure (ABM). These differ from N-gram, the previous corpus-based measure, in that the latter concentrates on the target information.

3.1. Sentence-Based Measure (SBM)

We assume that when we look at a bilingual corpus, i.e., a collection of model translations, if there is a sentence pair that is semantically similar to the pair of input sentence and its machine-generated translation, the machine-generated translation will probably be correct (Figure 2).

Getting back to the origin of this measure, Su [9] proposed automatic evaluation of translation quality by DP-matching of the machine generated translation and its reference translation. Yasuda et al. [17] extended the automatic evaluation to deal with variations inherent in natural language and to reduce the dependency of a single reference translation. Their method retrieves similar source sentences and makes a set of target counterparts of the similar sentences for reference in evaluation. Akiba [1] also extended Su's automatic evaluation to incorporate multiple DP-distances, including semantic ones, and harmonize the multiple distances by a decision tree learner. A combination of these two extensions comprises our proposal:

a) First, in advance, the source and target sentences in the bilingual corpus are converted to the **semantic category sequences** of *content words*. Thus, the corpus is converted to a table consisting of pairs of semantic category sequences.
b) Second, in run-time, the input sentence and machine-generated translation are converted **into semantic category sequences**.
c) Third, the similar source sequences are retrieved and the target sequences are compared with the translation sequence.
d) The minimum DP-distance is returned as the reliability value of the machine-generated translation.

With this reliability value, if the translation is incorrect, the value is large.

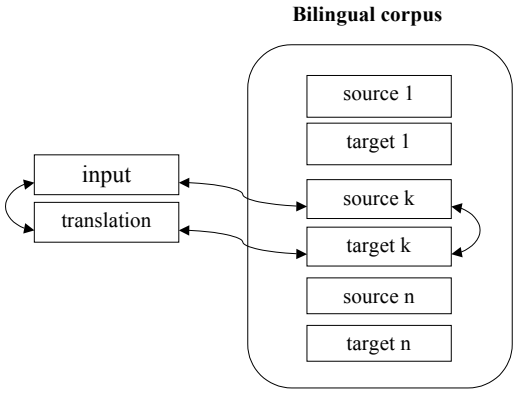

Figure 2 SBM uses bilingual corpus

3.2. Alignment-Based Measure (ABM)

In order to evaluate the translation, let us align an input sentence and its machine-generated translation. We assume that the better the obtained alignment is, the better the machine-generated translation.

a) **Align** the *content words* of input sentence and its machine-generated translation in an appropriate manner.
b) Compute the ratio of alignment according to **Tanimoto's coefficient.** If a source is aligned to a target word, they are regarded as identical in the calculation.

Suppose that E is the set of content words of the source sentences and that J is the set of content words of the target sentences, Tanimoto's coefficient is defined by the next equation.

$$Tanimoto = \frac{|E \cap J|}{|E \cup J|}$$

A larger Tanimoto's value indicates better alignment between the two.

4. EXPERIMENT

We conducted experiments on the effectiveness of the proposed reliability measures through selection of the best one from among multiple translations.

4.1. Conditions

4.1.1. Component machine translation systems

We prepared three different machine translation systems: TDMT [13], D3 [11] and SMT [19] for the travel conversation domain.

These are based on different paradigms, different development styles, and development periods. TDMT and D3 follow example-based approaches, but the former uses shorter templates handcrafted by experts while the latter uses longer templates automatically acquired from the corpus. SMT is a pilot system of a statistics-based approach that succeeded in translation between European languages. Development periods are shorter in the order of TDMT, D3 and SMT. This results in various differences for each input sentence, and the rank of translations changes sentence-by-sentence.

4.1.2. Bilingual Corpus and dictionaries

We built a collection of Japanese sentences and their English translations, which are usually found in phrasebooks for foreign tourists [14]. Because the translations were made sentence-by-sentence, the corpus was sentence-aligned at birth. We lemmatized and POS-tagged both the Japanese and English sentences by using our morphological analysis programs. The total sentence count was about 200 K. The statistics are summarized in Table 1.

Table 1 Corpus statistics

Sentences	204,108
Sentence Length	(J) 8.3
	(E) 6.1
Words	(J)1,689,449
	(E)1,235,747
Vocabulary	(J) 19,640
	(E) 15,374

A quality evaluation was done for a test set of 1,524 sentences selected randomly from the above corpus, and the remaining sentences were used for verification and learning. Translations of the test set by the three MT systems were evaluated and classified into 4 ranks by a professional translator according to the criteria of [13]. We also used a bilingual dictionary previously developed for TDMT. The size of the dictionary is 24,658 words. We also used thesauri previously developed for TDMT. The size of the Japanese thesaurus is 21,608 and that of the English thesaurus is 11,359.

4.2. Word alignment in the experiment

We developed a word-alignment program for ABM. It uses three kinds of lexical knowledge: (1) hand-made bilingual dictionary, (2) hand-made thesaurus, and (3) automatically acquired bilingual dictionary by cosine based word-alignment [10].

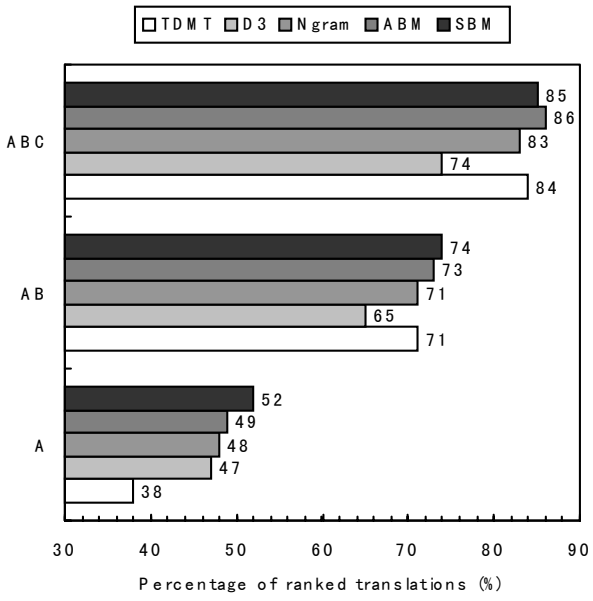

Figure 3 Component MTs and Three Selectors

The alignment is done in a cascade manner step-by-step based on the three kinds of lexical knowledge. The order reflects the reliability of the alignment based on the knowledge, so the later step does not change the alignments made by the previous steps.

4.3. Results

4.3.1. Selection accuracy

We simplified the problem to select the best one of the two translations produced by TDMT and D3.

As shown in Figure 3, the selector using the proposed ABM and SBM outperformed all component MTs in all ranks: A, A+B, A+B+C. Furthermore, they beat the conventional measure using N-gram.

Here are samples that were mistaken by N-gram and judged correctly by ABM and SBM. As pointed out in section 2.1, the source and target correspondences from the semantic point of view are not necessarily kept in state-of-the-art translation systems.

J2. ashita/no/heya/no/yoyaku/o/o-negai/shi/masu
{tomorrow/of/room/of/reservation/OBJ/hope/do/POLITE}
[D] Please make a reservation for room.
[A] I'd like to make a room reservation for tomorrow, please.

J3. zutuu/ga/shi/masu/asupirin/wa/ari/masu/ka
{headache/SUB/do/POLITE/aspirin/TOPIC/be/POLITE/QUESTION}
[C]I have a headache. Can you recommend an aspirin?
[A]I have a headache. Do you have any aspirin?

4.3.2. Problem and causes

Our measures often produced the same score for different translations. This occurred about 40% of the time. Half of these are not problematic because the two different translations deserve the same rank. The remaining half is problematic, so we should have discriminated the two translations. In the experiment, selection was decided by the majority.

This is caused by the granularity of our measures being too coarse, as illustrated in the next samples. J4 exemplifies the flaws of ABM in that ABM neglects word order. J5 exemplifies the flaws of SBM in that SBM does not distinguish the differences in level of lexical knowledge used for alignment.

J4. kousaten/wo/migi/ni/magari/nasai
{tomorrow/of/room/of/reservation/OBJ/hope/do/POLITE}
Turn at the crossing of the right.
Turn right at the crossing.

J5. hougaku/wo/senkou/shi/mashi/ta
{law/OBJ/major-in/do /POLITE/PAST}
I majored in mechanical engineering.
I majored in law.

"Law" is aligned to "hougaku" according to the bilingual dictionary, while "mechanical engineering" is aligned to "hougaku" according to thesauri because they share the semantic class of *learning*. Therefore, the current problem will be solved by refining the measures when we obtain the same score.

5. CONCLUSIONS

This paper proposed two reliability measures for translation quality that can be applied to any machine translation system. We conducted experiments on selecting the best one of multiple system outputs by using the proposed measures.

Our selector outperformed not only the component systems but also a conventional selector using N-gram, proving the effectiveness of our measures.

Future work includes (1) refining our measures and (2) investigating the integration of our and other[1] complementary measures.

[1] We have just devised a new measure [2] based on models of SMT and our experiment has demonstrated higher accuracy than that of two measures in this paper.

6. REFERENCES

[1] Akiba, Y., Imamura, K. and Sumita, E. 2001 Using multiple Edit Distances to automatically rank machine translation output, Proc. of MT-SUMMIT-VIII

[2] Akiba, Y., Watanabe, T. and Sumita, E. 2002 Using Language and Translation Models to Select the Best among Outputs from Multiple MT systems, Proc. of Coling

[3] Callison-Burch, C. and Flournoy, S. 2001 A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines, proc. of MT-SUMMIT-VIII

[4] Kaki, S., Yamada, S. and Sumita, E. 1999 Scoring Multiple Translations Using Character N-gram, Proc. of NLPRS:298-302

[5] Knight, K. and Chander, I. 1994 Automated Post-editing of Documents, Proc. of AAAI:779-784

[6] Ney, H., Och, F. J. and Vogel, S. 2000. Statistical Translation of Spoken Dialogues in the Vermobil System, Proc. of MSC2000: 69-74.

[7] Shimohata, M. and Sumita, E. "Automatic paraphrasing based on parallel corpus for normalization," LREC-2002.

[8] Somers, H. 1999 Review Article: Example-based Machine Translation, Journal of Machine Translation:113-157

[9] Su, K. -Y, Wu, M. -W. and Chang, J. –S. 1992 A new quantitative quality measure for machine translation systems, Proc. of Coling:433-439

[10] Sumita, E. 2000 Word alignment using matrix, Proc. of PRICAI:821

[11] Sumita, E. 2001 Example-based machine translation using DP-matching between word sequences, Proc. of DDMT workshop of 39th ACL: 1-8

[12] Sumita, E. 2002 Corpus-Centered Computation, Proc. of S2S workshop of 40th ACL

[13] Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K. and Shirai, S. 1999 Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach, Proc. of MT Summit VII:229-235

[14] Takezawa, T. et al. 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, Proc. of LREC-2002

[15] Tidhar, D. and Kuessner U. 2000. Learning to Select a Good Translation. Proc. of Coling-2000

[16] Yamamoto, K. 1999 Proofreading Generated Outputs: Automated Rule Acquisition and Application to Japanese-Chinese Machine Translation, Proc. of 18th ICCPOL:87-92

[17] Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M. 2001. An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus, Proc. of MT-SUMMIT-VIII

[18] Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M. 2002. Automatic Machine Translation Selection Scheme to Output the Best Result, Proc. of LREC

[19] Watanabe, T., Imamura, K. and Sumita, E. 2002 Statistical Machine Translation Based On Hierarchical Phrase Alignment, Proc. of TMI