Objective evaluation of second language learner's translation proficiency using statistical translation measures

Hajime Tsubaki[1], Keiji Yasuda[2], Hirofumi Yamamoto[2,3] and Yoshinori Sagisaka[1,2]

[1]GITI / Language and Speech Science Research Laboratories, Waseda University, Japan

[2]School of Science and Engineering, Kinki University, Japan

https://doi.org/10.36505/ExLing-2008/02/0055/000114

Abstract

For objective evaluation of the second language learner's translation proficiency, we tested two measures commonly used in statistical machine translation. One is word n-gram probability of a target language to measure likelihood of translated sentences as a target language. The other one is translation probabilities from source language sentences to target language sentences to measure translation accuracies between these sentences. The subjective proficiency scores of Japanese learners were compared with these objective measures extracted from English sentences translated by them. Statistical analysis showed no correlation of word n-gram probabilities but high correlation of translation probabilities which suggests the usefulness for objective evaluation of learner's translation proficiency.

Key words: objective proficiency evaluation, word n-gram, translation probability

Introduction

In language processing, quite a few studies have been carried out to automatically evaluate machine translation. Statistical measures such as NIST and BLEU have been extensively studied to emulate human's evaluation characteristics in comparison between man and machine (K. Papineni et al., 2002). These statistical objective measurements have also been effectively used in automatic evaluation of human learner's translation capabilities. In the evaluation of second language proficiency, these measures have been employed to evaluate proficiency by computing statistical differences between learner's sentences and native's ones by replacing an output from a translation system (Yasuda et al., 2003).

Though these studies have shown possibilities of objective evaluation for second language proficiency using comparative measures, they require correct answer sentences of a test set. It is quite laborious to prepare correct answers for every test sentences. To be free from this tedious data collection, we tried to use two measures, word n-gram probability and translation probability, employed in machine translation for learner's proficiency evaluation. If some of statistical translation measures are useful in the

evaluation of learner's sentences, we need not be bothered to prepare answer sentences for every test sets.

Measures for object evaluation

To estimate second language learner's proficiency from translated sentences, we need measures used by native raters. Considering their proficiency rating, we can find that they use multiple criteria such as (1) Word correspondences in translation, (2) Likelihood as an English expression, (3) Grammaticality, (4) Recoverability or seriousness of miss-translation and (5) Adequacy of corresponding target word selection. For objective evaluation, it is ideal to define a quantitative measure integrating all these criteria. However, in reality, it is not so easy to quantify what factors are relating how in subjective evaluation. It is difficult not only to list up all factors but also to prepare reasonable amount of learner's corpora to get reliable results. In this study, as a first step, we expect that two measures used in statistical translation, word n-gram probability and translation probability, can reflect the first two criteria.

As well known, in statistical machine translation from a Japanese sentence j to an English sentence e, English sentence *e* that maximizes P(e|j) in all translation candidates is selected by using two statistics P(e) and P(j|e) as expressed in the following equation.

$$e = \underset{\text{all candidate}}{\operatorname{argmax}} P(e|j) = \underset{\text{all candidate}}{\operatorname{argmax}} P(j|e) \cdot P(e)$$

where P(j|e) stands for translation probability and P(e) corresponds to occurrence probability of English sentence e. Two measures that we use for our analysis correspond to these two probabilities.

The translation probability P(j|e) is calculated using IBM Model 1, word-based translation model (Peter E Brown et al., 1993). P(j|e) is obtained by word-to-word translation probabilities between the English and Japanese sentence. On the other hand, English sentence probability P(e) is approximated by word 3-grams probabilities $P(w_i|w_{i-2}, w_{i-1})$ as follows.

$$P(e) \approx \Pi P(w_i|w_{i-2}, w_{i-1})$$

In the evaluation, we apply the above calculation formula by considering an English sentence translated by a learner as a translation candidate in statistical translation.

Evaluation experiment using two statistical measures

We calculated English word n-gram probability and translation probability from a Japanese sentence to an English one to evaluate their effectiveness for sentence accuracy and learner's proficiency. Using these probabilities, we got correlation scores between these measures and subjective scores for test

set sentences. Finally, we calculated the correlation between learner's proficiencies and objective scores using an effective probability.

Experimental setup

For the evaluation experiment, we employed a sentence set consisting of Basic Traveler's Expression Corpus (BTEC) (Takezawa et al., 2002) and learner's corpus. Translation probability P(j|e) and word 3-grams probability $P(w_i|w_{i-2}, w_{i-1})$ were calculated using the BTEC. The learner's corpus consists of 473 source Japanese sentences translated by 21 learners with different English proficiencies and evaluated in five scales by a Japanese-English bilingual rater based on evaluation criteria (S:Native, A:Good, B:Fair, C:Acceptable and D:Nonsense).

Experimental results and discussions

Figure 1 shows averages and standard deviations of (a) translation probabilities and (b) word n-gram probabilities over all sentences belonging to each subjective scoring category from D (lowest) to S (highest). As Figure 1 shows, the translation probability average increases as subjective score becomes high. On the other hand, the word n-gram probabilities show no correlation between subjective scores. The positive correlation in translation probability indicates its usefulness in the objective evaluation of sentences.

To confirm the validity of translation probability for the objective evaluation of learner's proficiency, we calculated correlation between the averages of translation probabilities for each learner and the learner's TOEIC scores. The correlation was turned out to be 0.287. By analyzing the data, we found that this low correlation results from lower correlations in short sentences. To quantify the effects of sentence length, we measured the correlations of subgroups divided by their sentence length. As shown in Table 1, we could find the increase of correlations between the averages of translation probabilities and TOEIC scores in proportion to sentence length.

These results suggest that the translation probability is useful for objective evaluation of learner's proficiency. We need further studies for more efficient use of it by taking test sentence length or sentence complexities into account. While, the word n-gram possibility turned out to be of no use even for sentence evaluation. This result is quite different from the expectations from related previous works (Yasuda et al., 2003). As word statistics were differently used in this work, the word n-gram possibility might not show any differences. Moreover, all learners tend to translate using word sequences that they are familiar with, their occurrence possibilities may not vary so much according to their proficiencies. We should employ lexical information such as word difficulty ranking or expressions directly reflecting proficiencies.

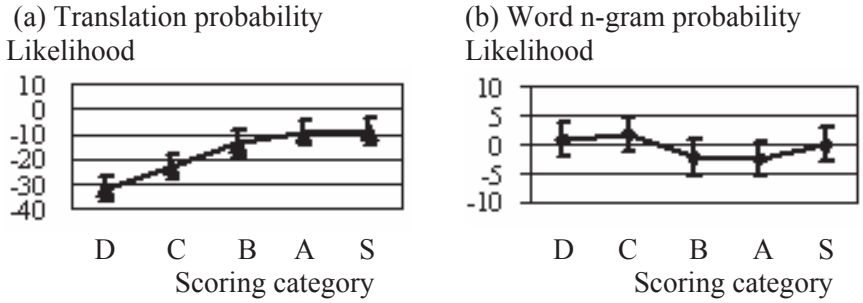

Figure1:Average and standard deviation of translation probabilities and word n-gram probabilities over all sentences belonging to each subjective scoring category.

Table1. Increase of the correlation score between translation probabilities and learner's TOEIC scores to the sentence length in words (Japanese)

Japanese sentence length	1 ~ 5	6 ~ 10	11 ~ 15	16 ~ 20	21 ~
Correlation score	0.240	0.300	0.351	0.404	0.467

Conclusions

To obtain effective measures for objective evaluation of learner's second language proficiency without being bothered by tedious correct data collection, we have tested the availability of translation probability and word n-gram probability. The analysis experiment showed the usefulness of translation probability for sentence translation evaluation and learner's TOEIC scores. We also found that further specification of measure would increase its effectiveness. We will continue to find parameterizations of other measures that we have not yet used together with effective use of learner's data by themselves.

References

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. Proc. of ACL2002, pp. 311-318.

Keiji Yasuda, Eiichiro Sumita, Seiichi Yamamoto, Masuzo Yanagida, Kikuo Maekawa, and Fumiaki Sugaya. 2003. A Proposal for Automatically Gauging of English Language Proficiency. IPSJ SIG Technical Report, Vol.2003-NL-155: 65-70.

Peter E Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19(2):263-311

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. LREC2002, pp.147-152.