A Method of Automatically Evaluating Machine Translations Using a Word-alignment-based Classifier

Katsunori Kotani
Kansai Gaidai University/NICT
3-5 Hikaridai, Seika-cho,
Soraku-gun, Kyoto, Japan
kat@khn.nict.go.jp

Takehiko Yoshimi
Ryukoku University/NICT
3-5 Hikaridai, Seika-cho,
Soraku-gun, Kyoto, Japan

Takeshi Kutsumi
Sharp Corporation
492 Minosho-cho, Yamato-koriyama, Nara, Japan

Ichiko Sata
Sharp Corporation
492 Minosho-cho, Yamatokoriyama, Nara,
Japan

Hitoshi Isahara
National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto,
Japan

Abstract

Constructing a classifier that distinguishes machine translations from human translations is a promising approach to automatically evaluating machine-translated sentences. We developed a classifier with this approach that distinguishes translations based on word-alignment distributions between source sentences and human/machine translations. We used Support Vector Machines as machine-learning algorithms for this classifier. Our experimental results revealed that our method of evaluation had a weak correlation with human evaluations. We further found that our method outperformed well-known automatic-evaluation metrics with respect to correlation with the manual evaluation, and that it could identify the qualitative characteristics of machine translations, which greatly help improve their quality.

1 Introduction

Previous research has proposed various automatic methods of evaluating machine-generated translations (MTs). Some methods have examined the similarity of MTs to human-generated translations (HTs), i.e., BLEU (Papineni et al. 2001), NIST (Doddington 2002), METEOR (Banerjee & Alon 2005), Kulesza & Shieber (2004), Paul et al. (2007), and Blatz et al. (2004). These methods would be rather expensive due to the need to prepare multiple-reference HTs for evaluation. To resolve this problem, Corston-Oliver et al. (2001) and Gammon et al. (2005) proposed methods of evaluation, which did not employ multiple reference HTs in evaluating MTs.[1] Instead of evaluating MTs by comparing them with HTs, evaluation was carried out with a machine-learning algorithm that classified MTs either into "good" or "bad" translations. A "good" translation is a translation that is indistinguishable from HTs, whereas a "bad" translation is a translation that is judged to be an MT. Although this method of classification might require reference HTs to construct a classifier as training data, it does not need any reference HTs for evaluation. Hence, once a classifier is constructed, this method can be applied to any translations without reference HTs. This is an advantage of classifier-based evaluation methods.

This new method also reveals what sorts of errors are involved in MTs, while others such as BLEU (Papineni et al. 2001) cannot, as Corston-Oliver et al. (2001) suggested. The primary goal of BLEU (Papineni et al. 2001) was to determine the superiority of translation systems, and hence, the method outputs numerical values in terms of BLEU scores. When one tries to improve a translation system, it is necessary to identify the problems with it. On these grounds, we surmised that a clas-

[1] Albrecht & Hwa (2007) also proposed MT evaluation metrics without using reference HTs. Their method employed the regression-trained metric.

sifier-based scheme would be a promising approach to evaluating MTs.

Although source sentences need to be referred to in order to evaluate the adequacy of MTs, these previous methods have only examined the linguistic properties of MTs but not those of source sentences. Hence, they have focused on the fluency of translation but not on the adequacy of translation. Adequacy is defined as to what extent a translated sentence conveys the meaning of the original sentence. Fluency is defined as the well-formedness of a translated sentence, which can be evaluated independently of adequacy.

This paper discusses our examination of a classifier, which can evaluate MTs from both viewpoints of fluency and adequacy. In evaluating translations from English to Japanese, for instance, not only the translation fluency but also its adequacy should be carefully assessed, because translations between these languages involve greater linguistic problems than those between European languages, e.g., English and French. European languages belong to the same language class, whereas English and Japanese do not. Thus, English and Japanese vary greatly with respect to various linguistic properties such as anaphoric systems (see Section 3.3).

This linguistic divergence makes evaluations of adequacy significant for MTs of English into Japanese. We propose employing a classification feature that reveals the linguistic correspondences between source sentences and translations to evaluate adequacy with the classification method. Incidentally, unlike reference translations, source sentences are necessary to obtain MTs. Thus, we constructed a classifier that would distinguish translations based on word-alignment distributions between source sentences and translations, assuming that the word-alignment distributions exhibited linguistic correspondences between these source sentences and translations. We then assessed our method by comparing its evaluation results with those of human evaluations.

2 Method of Machine Learning

Our method uses Support Vector Machines (SVMs), which are well known learning algorithms that have high degrees of generalization. We used SVMs to build a classifier based on word-alignment distributions as machine-learning features.

Our method employs parallel corpora to construct the classifier and requires neither manually labeled training examples (unlike Albrecht) nor multiple reference translations to evaluate new sentences. Due to these properties, our method should be a relatively inexpensive but effective automatic evaluation metric.

2.1 Evaluation Metric Obtained by SVMs

SVMs are learning algorithms based on maximum margin strategy (Vapnik 1998). We train an SVM classifier by taking HTs as positive training examples and MTs as negative. Consequently, the SVMs produce a hyperplane that separates the examples. As Kulesza & Shieber (2004) noted, the distance between the separating hyperplane and a test example can serve as an evaluation score. Based on this idea, our classifier not only distinguishes the MTs from HTs but also evaluates the MTs with this metric.

2.2 Features

As we noted in Section 1, word-alignment distribution should constitute classification features examining translation adequacy. We further presumed that word-alignment distribution could also be used to examine translation fluency.

Good, natural translations differ from poor, unnatural translations such as word-for-word translations, because superior translations involve various translation techniques. For instance, there is a technique for translating the English nominal modifier “some” into a Japanese existential construction, as in (1b) below. The meaning of the English nominal modifier is conveyed in the existential verb *i-ta* “existed”. The translation of (1a) without this technique, i.e., by word-for-word translation, is presented in (1c), where the English nominal modifier “some” is translated into the Japanese nominal modifier *ikuraka-no* “some”. Translation (1c) is perfectly grammatical but less natural than (1b). Actually, sentence (1c) was obtained with a state-of-the-art MT system. If this translation technique were implemented on this system, the system would produce a more natural sentence.

As example (1) illustrates, because MTs are literally translated, they often sound unnatural. Therefore, we decided to compare MTs and HTs regarding the degree of word-for-word translation. To identify word-for-word translation, we used the word-alignment distribution between source sentences and translations, i.e., MTs or HTs, because literally translated words should be more easily aligned than non-literally translated words. Literal translations maintain lexical features such as parts of speech, as can be seen in (1). By contrast, non-literal translations usually lack parallel lexical features.

```
(1)
a.   Some students came.
b.   Ki-ta  gakusei-mo i-ta
     come-PST student-also exist-PST
     “Some students came”.
c.   Ikuraka-no gakusei-wa ki-ta
     some-GEN  student-TOP come-PST
     “Several students came”.
(GEN: Genitive case marker,
PST: Past tense marker, TOP:
Topic marker)
```

Figure 1. Translation Example (1)

Let us illustrate the difference in alignment distribution between MTs and HTs.

```
(2)
a.   Today, the sun is shining.
b.   Kyoo taiyoo-wa  kagayai-teiru
     today the-sun-TOP shine-BE-ING
     “Today the sun is shining”.
c.   Kyoo-wa  seiten-da
     today-TOP fine-BE
     “It’s fine today”.
(TOP: Topic marker, BE: Copular
verb, ING: Gerundive verb
form)
```

Figure 2. Translation Example (2)

Sentence (2a) below is a source sentence both for the word-for-word translation in (2b), i.e., MT, and the natural translation in (2c), i.e., HT. Table 1 lists the word-alignment distribution attained with our alignment tool. In Tables 1 and 2, “align(A, B)” means that an English word “A” and a Japanese word “B” compose an aligned pair, “non-align_eng(C)” means that an English word “C” remains unaligned, and “non-align_jpn(D)” means that a Japanese word “D” remains unaligned. From the alignment distribution in Tables 1 and 2, we see that the rate of alignment and non-alignment varies between HTs and MTs. That is, non-aligned words often appear in HTs, and more aligned pairs are observed in MTs. Thus, non-aligned words should exhibit HT-likeness, while aligned pairs should exhibit MT-likeness. We constructed a classifier using these aligned pairs and non-aligned words in Tables 1 and 2 as classification features. Since word-alignment properties reveal the lexical correspondences between a source sentence and its counterpart, our classifier can take adequacy into account.

Table 1. Alignment Distribution of MTs

MT (2b)
align(today, kyoo [today])
align(is, teiru [BE-ING])
align(sun, taiyoo [sun])
align(shining, kagayai [shine])
nonalign_jpn(wa [TOP])
nonalign_eng(the)

Table 2. Alignment Distribution of HTs

HT (2c)
align(today, kyoo-wa [today-TOP])
align(is, da [BE])
nonalign_jpn(seiten [fine])
nonalign_eng(the)
nonalign_eng(sun)
nonalign_eng(shining)

3 Experiments

This section describes the design and results of our experiment, and discusses our findings.

3.1 Design

A parallel corpus was prepared for constructing classifiers in the experiment. The corpus consisted of Reuters’ news articles in English and their Japanese translations (Utiyama & Isahara 2003). Since some source sentences and translations appeared repeatedly in our corpus, we deleted these repetitions. The MTs for this corpus were obtained with a commercially available MT system. Word-alignment distributions between the source sentences and the MTs and HTs were obtained with an experimental word-alignment tool.[2] A total of

[2] Experiments with a free alignment tool (Och & Ney 2003) have yet to be done.

258,000 examples were obtained (129,000 HT-alignment examples and 129,000 MT-alignment examples).

We randomly chose 44 sentences from this corpus for a preliminary evaluation of our method.[3] These sentences were assessed by three human evaluators, who had been involved in developing MT systems (not the authors). The evaluators assessed both the adequacy and fluency of MTs, and scored them on a scale from 1 to 4. (See Section 1 for the definitions of adequacy and fluency.)

Machine learning was carried out with an SVM algorithm implemented on the TinySVM software.[4] The linear was taken as a type of kernel function, and the other settings were taken as default settings.

We first appraised the accuracy of classification with our classifier in this experiment. Then, we investigated the correlation between the human-assessment results obtained by our three evaluators to determine the upper bounds for our classification-based method. Finally, we investigated and tested its validity by examining how well the scores computed by the SVMs correlated with the adequacy and fluency scores awarded by the human evaluators.

3.2 Results

Before reporting the experimental results, let us briefly confirm the word-alignment distributions in MTs and HTs. As Table 3 shows, the number of aligned pairs constituted 35% of alignment distributions in MTs. By contrast, the aligned pairs made up 24% in HTs. In Table 3, the number refers to the sum of the aligned pairs and non-aligned words between the 129,000 source sentences and the MTs/HTs. Thus, MTs contain more aligned pairs than HTs. We tested the differences in alignment distributions between HTs (control sample) and MTs with a Fisher exact test. The results revealed that the alignment rate for MTs was significantly greater than that for HTs ($p<0.05$). Based on these results, we concluded that MTs and HTs differed with respect to word-alignment distributions.

Table 3. Alignment Distributions

	N	Aligned pairs (%)	Non-aligned words (%)	Alignment rate (%)
MT	521102	35.7	64.3	55.5
HT	568259	24.1	75.9	31.7

Next, we examined the robustness of our method for machine translation systems by comparing the classification accuracy of three commercially available state-of-the-art translation systems in a five-fold cross validation test. Our method of classification yielded high classification accuracy (98.7, 99.7%, and 99.8%). From these results, we concluded that our method is robust for MT systems.

Now, let us return to the results from the experiment. First, we examined the classification accuracy of our classifier. Its accuracy was obtained through the five-fold cross validation test. Our method of classification achieved a high accuracy of 98.7%. It is difficult to find benchmark methods to compare with our classifier, because previous methods often require multiple reference translations or manually labeled training examples. Since the previous studies used syntactic properties to construct classifiers (Corston-Oliver et al. 2001, Gamon et al. 2005, Mutton et al. 2007), we decided to compare our alignment-distribution-based classifier with a classifier based on syntactic properties, i.e., dependency relations. Although this comparison was not that rigorous, we believe it suggested that our method was valid. HTs and MTs were parsed with the CaboCha parser (Kubo & Matsumoto 2002), and the dependency pairs of a modifier and a modified phrase were used as classification features. This baseline method achieved an accuracy of 83.1%. Our proposed method outperformed the baseline, exhibiting a superiority of 18.8%. Based on these results, we concluded that our word-alignment-based classifier more accurately distinguishes MTs and HTs than a dependency-relation-based classifier.

We next checked the correlation of assessment results between the three human evaluators (I-III). The results for both adequacy and fluency exhibited strong correlations as listed in Table 4. The correlation coefficients for adequacy evaluation varied from .68 to .76, and those for fluency evaluation ranged from .40 to .61. We determined

[3] The number of test sentences should be increased in future experiments to enable more rigorous evaluations of our method. We are now preparing a larger-scale experiment.

[4] The packaging tool is available at the following URL: http://chasen.org/~taku/software/TinySVM/

the upper bounds for our classifier as the mean values of human evaluation. That is, the bound for adequacy evaluation was .73, the bound for fluency evaluation was .53, and the bound for the entire evaluation was .74. The entire evaluation was derived by summing up both adequacy and fluency evaluation scores.

Table 4. Correlation of Human Evaluation Results

	I-II	I-III	II-III	Mean
Adequacy	.76	.74	.68	.73
Fluency	.58	.40	.61	.53
Entire	.76	.70	.75	.74

Finally, we moved on to evaluating the performance of our method. We examined to what extent our classifier-based evaluation results were correlated with the human-evaluation results. The correlations were examined at the sentence level. The MT sentences were evaluated with our method using a score provided by the SVM classifier as described in Section 2.1. The human evaluation consisted of three types of evaluation scores: (i) adequacy, (ii) fluency, and (iii) entire. We assessed our evaluation method (W-A classifier) by comparing it with human evaluations. In addition, we evaluated three other methods: (i) a dependency-based classifier (D-classifier), (ii) NIST (Doddington 2002), (iii), and METEOR (Banerjee & Alon 2005). The correlations were assessed in terms of Spearman's rank-correlation coefficient.

Table 5. Correlation of Automatic-evaluation Results and Human-evaluation Results

	Adequacy	Fluency	Entire
W-A classifier	.44	.43	.47
D-classifier	.33	.37	.37
NIST	.40	.45	.46
METEOR	.20	.19	.20

Table 5 lists the correlation coefficients. In obtaining the evaluation results for NIST (Doddington 2002) and METEOR (Banerjee & Alon 2005), we used HTs of the parallel corpus as reference translations.

3.3 Discussion

Our classification-based method of evaluation, which employed word-alignment distributions as learning features, exhibited a weak correlation with the human-evaluation results for adequacy, fluency, and the entire evaluation, as listed in Table 5. Our method did not surpass the upper bound coefficients, i.e., the mean correlation coefficients between the human-evaluation results in Table 4.

Compared with the other three automatic methods, our classifier outperformed the D-classifier and METEOR (Banerjee & Alon 2005) in the three evaluation criteria, and our method achieved similar results to NIST (Doddington 2002). Our method had a lower correlation coefficient with human fluency evaluation than NIST (Doddington 2002), but it outperformed NIST (Doddington 2002) with respect to adequacy and the entire evaluation. The D-classifier-based method of evaluation did not achieve as high a correlation as NIST (Doddington 2002). From these results, we assumed that our method was tenable as an automatic method of evaluation without the use of reference translations. In addition, our method seems to account for evaluations of adequacy as we assumed that these need to be examined with features covering linguistic correspondences between source sentences and translations, i.e., word alignments (as discussed in Section 2.2). The correlation coefficient of adequacy evaluation for the D-classifier-based evaluation was lower than that of fluency evaluation. By contrast, the adequacy evaluation achieved a higher correlation than the fluency evaluation in the W-A-classifier-based evaluation. This suggests that the W-A-classifier-based evaluation appropriately assessed the adequacy evaluation. We intend to test and verify this conclusion in future studies.

We further appraised the experimental results by comparing them for our method and human evaluation. We consequently found that while fluency evaluation decreased in human evaluation, automatic-evaluation methods (including ours) did not exhibit such drops. All the automatic-evaluation methods exhibited similar correlations between adequacy and fluency evaluations. Hence, unlike human evaluation, automatic evaluation seems stable for evaluating fluency. This constitutes one advantage of automatic evaluation.

Using word-alignment distribution as classification features, we can construct three types of classifiers: (i) a classifier based on aligned pairs (AL), (ii) a classifier based on non-aligned words (n-AL), and (iii) a classifier based on both aligned pairs and non-aligned words (AL & n-AL). We compared the evaluation accuracy with these classifiers by comparing it with human evaluation. As listed in Table 6, the classifier using both aligned and non-aligned words achieved the highest correlation. Hence, this led us to employ both aligned and non-aligned distribution as classification features.

Table 6. Correlation of Classifier-evaluation and Human-evaluation Results

	Adequacy	Fluency	Entire
AL	.28	.32	.33
n-AL	.28	.27	.28
AL & n-AL	.44	.43	.47

In addition, our method could reveal problems with MT systems by enabling weights given to all features in training the SVM classifier to be assessed. The weight of a feature indicates its MT-likeness or HT-likeness with our method. The MT/HT-like properties are proportional to the absolute value of the weight.

Through investigating the weights of features, we found that well-known translation problems in MTs could be detected. As Yoshimi (2001) noted, the translation of English pronouns into non-pronominal Japanese expressions is an MT problem that needs to be resolved. This arises from the linguistic difference between English and Japanese. English is a language that frequently uses pronouns, whereas Japanese uses fewer pronouns. In investigating the weights, we found aligned English pronouns for MT-likeness features and non-aligned English pronouns for HT-likeness features.

Table 7. Weights for HT-like Features

Rank	HT-like	Weight
1	**nonalign_jpn(doo [the same])**	**1.134**
2	nonalign_eng(just)	0.884
3	**nonalign_jpn(doo-si [the same person])**	**0.846**
4	nonalign_jpn(kono [this])	0.805
5	nonalign_jpn(akiraka [clear])	0.727

Table 8. Weights for MT-like Features

Rank	MT-like	Weight
1	nonalign_jpn(paasento [percent])	-0.982
2	**align(and, sosite [and])**	**-0.915**
3	Align(delay, okure [delay])	-0.874
4	**align(and, oyobi [and])**	**-0.796**
5	nonalign_jpn(u [?])	-0.780

Tables 7 and 8 list the five most HT-like features and MT-like features, respectively. As we can see from Table 7, HTs involve "non-align_jpn(doo [the same])" and "non-align_jpn(doo-si [the same person])". These expressions remained non-aligned due to the application of a translation technique to HTs. Here, the meaning of an English pronoun seems to be conveyed with a non-pronominal suffix, "doo- [the same]". Based on how the features are weighted, we can see that this translation technique can be applied to HTs but not to MTs. This is illustrated by example (3). Here, the English pronoun "he" is translated into the Japanese pronoun "kare [he]" in MT (3b). In HT (3c), the English pronoun "he" is translated into "doo-si [the same person]", which conveys an anaphoric meaning more naturally than a pronoun in this context.

(3)
a. He said the policy would increase textile exports both in terms of value and quantity.
b. kare-wa itt-ta. Sono-hooosin-wa
he-TOP say-PST this policy-TOP
kati-no-aru-kikan-ni sosite mata
value-GEN-exist-span-DAT and also
ryoo-de senni-no yusyutu-wo
quantitiy-DAT textile-GEN exports-ACC
zooka-suru-de-aroo-to
increase-will-COMP
c. doo-si-wa sin-booeki-
the-same-person-TOP new-export-
seisaku-ga doonyuu-sareru-to
policy-NOM introduce-PASS-COMP
senni-yusyutu-wa kakaku-to-
textile-exports-TOP value-and-
ryoo-no ryoomen-de
quantities-GEN both-side-DAT
zooka-suru-to katat-ta
increase-COMP tell-PST
(TOP: Topic marker, PST: Past tense marker, GEN: Genitive case marker, DAT: Dative case marker, ACC: Accusative case marker, COMP: Complementizer, NOM: Nominative case marker, PASS: Passive marker)

Figure 3. Translation Example (3)

In addition to translating pronouns, we found that MTs and HTs differed in translating coordinating conjunctions. The English conjunction “and” can conjoin any categorial phrases such as noun phrases, verb phrases, and sentences. Japanese has both a categorially restricted free conjunction, i.e., “sosite [and]” and a restricted conjunction, i.e., “-to [and]”. The latter conjunction can only conjoin nominals. Thus, conjunctions constitute another linguistic discrepancy between Japanese and English. As Fujita (2000) suggests, the translation of the English conjunction “and” into Japanese conjunctive expressions is a translation problem that needs to be resolved. HTs seem to apply another translation rule to conjunctions. While HTs have no alignment features concerning conjunctions, MTs involve aligned pairs for conjunctive expressions, i.e., “align(and, sosiste [and])” and “align(and, oyobi [and])”, as listed in Table 8. This difference in translating conjunctions is also illustrated in example (3). In MT (3b), the English conjunction “and” is translated into “sosite [and]”, while a conjunction is translated into the conjunction suffix “-to” in HT (3c). Noun phrases are more naturally conjoined with the conjunction “-to [and]” than the other conjunction “sosite [and]”.

4 Conclusion

We proposed an automatic method of evaluating MTs, which does not employ reference translations for evaluation of new sentences. Our evaluation metric classifies the results of MTs into either “good” translations (HTs) or “bad” translations (MTs). The classifier was constructed based on the word-alignment relations between source sentences and HTs/MTs, assuming that the alignment distribution reflected MT-likeness and HT-likeness. The classification accuracy in our experiment was 98.7%. We found that this classification-based method of evaluation exhibited a weak correlation with human-evaluation results and that it was more highly correlated with human evaluations than NIST (Doddington 2002) or METOR (Banerjee 2005) metrics. Our examination of how features were weighted revealed problems that studies on MTs should contend with, e.g., translation anaphoric expressions and conjunctive expressions. Our method, which employs parallel corpora, is relatively inexpensive but is an effective automatic evaluation metric.

This paper leaves several problems unsolved. First, we must examine to what extent the alignment features account for the difference between MTs and HTs. Second, we plan to investigate and test the validity of the new method in more detail by comparing our evaluation results with the more extended results attained by human evaluators.

References

Albrecht, J. S. and R. Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL07)*. 296–303.

Banerjee, S. and L. Alon. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL05)*. 65–72.

Blatz, J., E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. *Final Report, JHU /CLSP Summer Workshop*.

Corston-Oliver, S., M. Gamon, and C. Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL01)*. 148–155.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd Human Language Technologies Conference (HLT02)*. 128–132.

Fujita, N. 2000. *Nihongo-bunpoo*. Aruku, Tokyo.

Gamon, M., A. Aue and M. Smets. 2005. Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th the European Association for Machine Translation Conference (EAMT05)*. 103–111.

Kubo, T. and Y. Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL 2002)*. 63–69.

Kulesza, A. and S. M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI04)*. 75–84.

Mutton, A., M. Dras, S. Wan, and R. Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of* *the Association for Computational Linguistics (ACL01)*. 344–351.

Och, F. J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29 (1): 19–51.

Papineni, K. A., S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. *Technical Report RC22176 (W0109–022)*, IBM Research Division, Thomas J. Watson Research Center.

Paul, M., A. Finch, and E. Sumita. 2007. Reducing human assessment of machine translation quality to binary classifiers. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*. 154–162.

Utiyama, M. and H. Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*. 72–79.

Vapnik, V. 1998. *Statistical Learning Theory*. Wiley-Interscience, New York.

Yoshimi, T. 2001. Improvement of translation of pronouns in an English-to-Japanese MT system. *Journal of Natural Language Processing* 8 (3): 87–106.