Using Multiple Edit Distances to Automatically Rank Machine Translation Output

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seikacho Soraku-gun
Kyoto 619-0288 Japan
{yakiba, kimamura, sumita}@slt.atr.co.jp

Abstract

This paper addresses the challenging problem of automatically evaluating output from machine translation (MT) systems in order to support the developers of these systems. Conventional approaches to the problem include methods that automatically assign a rank such as A, B, C, or D to MT output according to a single edit distance between this output and a correct translation example. The single edit distance can be differently designed, but changing its design makes assigning a certain rank more accurate, but another rank less accurate. This inhibits improving accuracy of rank assignment. To overcome this obstacle, this paper proposes an automatic ranking method that, by using multiple edit distances, encodes machine-translated sentences with a rank assigned by humans into multi-dimensional vectors from which a classifier of ranks is learned in the form of a decision tree (DT). The proposed method assigns a rank to MT output through the learned DT. The proposed method is evaluated using transcribed texts of real conversations in the travel arrangement domain. Experimental results show that the proposed method is more accurate than the single-edit-distance-based ranking methods, in both closed and open tests. Moreover, the proposed method could estimate MT quality within 3% error in some cases.

Keywords: automatic evaluation, machine translation system, multiple edit distances, decision trees, machine learning

1. Introduction

ATR has been developing the transfer-driven machine translation system, TDMT (Furuse et al., 1995; Furuse & Iida 1996; Sumita et al., 1999), which is used as a subsystem in a multi-lingual speech translation system called ATR's multilingual automatic translation system for information exchange, ATR-MATRIX (Takezawa et al., 1999; Yamamoto, 2000). Both TDMT and ATR-MATRIX are designed for the travel arrangement domain.

TDMT have been subjectively evaluated with four ranks: A, B, C, and D (Sumita et al., 1999). The four ranks are defined as follows: (A) Perfect: no problems in both information and grammar; (B) Fair: easy-to-understand, with either some unimportant information missing or flawed grammar; (C) Acceptable: broken, but understandable with effort; (D) Nonsense: important information has been translated incorrectly. Machine-translated sentences have been manually ranked by native speakers of target languages.

Such subjective evaluation by ranking, however, is taxing on both time and resources (King, 1996). The developers of TDMT would like to evaluate their MT system under development more frequently; therefore if automatic evaluation methods are inexpensive, fast, sufficiently accurate for them to assess whether or not the current version of their MT system is improved, then these automatic evaluation methods will prove beneficial.

Conventional approaches to automatic evaluation include methods (Thompson, 1991; Su, 1992; Takezawa et al., 1999; Sugaya et al., 1999; Yasuda et al., 2000; Yasuda et al., 2001) that automatically assign one of several ranks (Sumita et al., 1999; Nagao & Tsujii, 1985) such as A, B, C, and D to MT output according to a single edit distance between an MT output and a correct translation example.

The single edit distance can be differently designed, but changing its design makes assigning a certain rank more accurate, but another rank less accurate. For examples, ED*i* (*i* = 1, 5, 9, or 13) in Figure 1 differ from each other in its design. For ED1, the combination of edit operators: either Insertion, Deletion, or Replacement are applied to some of all words. Each ED*i* (*i* = 1, 5, 9, or 13) can be placed in the order of the correct acceptance ratio of a rank in four ways. ED1 has, respectively, the best, the third best, the worst, and the best correct acceptance ratio of A, B, C and D. This phenomenon inhibits the accuracy of rank assignment from being improved.

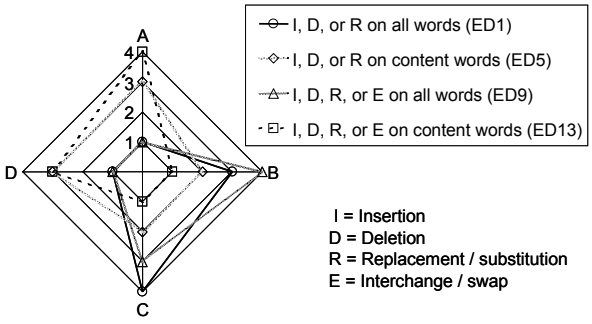

Figure 1: The order of ED*i* (*i* = 1, 5, 9, or 13) in the acceptance ratio of each rank estimated by the existing edit distance on all or content words (Winger, 1974; Lowrance, 1975). Refer notation of ED*i* (*i* = 1, 5, 9, or 13) to Section 2.2.

To overcome the inhibition, this paper proposes an automatic ranking method that, by using multiple edit distances, encodes machine-translated sentences with a rank assigned by humans into multi-dimensional vectors from which a classifier of ranks is learned in the form of a decision tree. The proposed method assigns a rank to MT output through the learned DT.

The proposed method is evaluated using transcribed texts of real conversations in the travel arrangement domain. Experimental results show that the proposed method is more accurate than the single-edit-distance-

based ranking methods, in both closed and open tests. The proposed method has the potential to estimate the quality of TDMT within 3% error in some cases.

The next section proposes our method. Experimental results are shown and a discussion is provided in Section 3. Finally, our conclusion is presented in Section 4.

2. Proposed Method

This section describes the proposed automatic ranking method by using multiple edit distances. The proposed method is based on two kind of information: (a) machine-translated sentences that are ranked by three or more human evaluators, and (b) plural sentences correctly translated by humans. The reason why the former information is used is that the proposed method is expected to assign a averaged rank to MT output. On the other hand, as Thompson (1991) and King (1996) state, it is in nature of translation that for given text, potentially many translations would all be equally acceptable. The proposed method, therefore, uses the latter information. The latter information is called multiple standards by Thompson (1991).

2.1. Outline of Our Method

The outline of the proposed method is as follows:

(1) Label each machine-translated sentence by the majority rank. For example, if the machine-translated sentence is ranked as "A" by two evaluators and as "C" by another evaluator, then the machine-translated sentence is labeled "A". If the machine-translated sentence is ranked as "A" by one evaluator, as "B" by another evaluator, and as "C" by another evaluator, then the machine-translated sentence is labeled "B".

(2) Encode each machine-translated sentence into a sixteen dimensional vector. The value of the *i*th element is calculated in the almost same way to (Thompson, 1991). The difference is the editing unit as explained in the next section. That is, the *i*th element is filled with the minimum value of edit distance ED*i*, listed in Table 2, between the machine-translated sentence to be encoded and a human-translated sentence. For example, as in Figure 2, if the number of different human-translated sentences whose source sentences are the same as that of the machine-translated sentence to be encoded is three, ED1 is calculated for each of the three pairs:{(T), (H1)}, {(T), (H2)} and {(T), (H3)}; therefore, the 1st element is filled with the minimum value in the three resulting values from ED1. The detail of each edit distance ED*i* (*i*=1,…,16) in Table 2 is explained in the next section.

(3) Learn a decision tree from the vectors.

(4) Assign a rank to MT output by using the learned decision tree.

The data flow of the proposed method is illustrated in Figure 3. Learning phase consists of the above three steps: (1), (2), and (3). Evaluation phase consists of the above two steps: (2) and (4).

(S) *They are a couple coming to Spain for sightseeing.*

(T) *Φ kankou-no supein-ni kuru kappuru desu.*
THEY SIGHTSEEING-OF SPAIN-TO COMING COUPLE ARE
' They are a couple coming to Spain of sightseeing.'

(H1) *karera-wa supein-he kankou-ni kuru hutari desu.*
THEY SPAIN-TO SIGHTSEEING-FOR COMING COUPLE ARE
' They are a couple coming to Spain for sightseeing. '

(H2) *Φ kankou-shi-ni supein-ni kuru kappuru desu.*
THEY SIGHTSEEING-DO-TO SPAIN-IN COMING COUPLE ARE
' They are a couple coming to sightsee in Spain. '

(H3) *Φ kankou-de supein-ni otozureru hutari desu.*
THEY SIGHTSEEING-FOR SPAIN-loc VISITING COUPLE ARE
' They are a couple visiting Spain for sightseeing. '

Figure 2: An example of both output (T) from English-to-Japanese MT system and human-translated sentences: (H1), (H2), and (H3). Each translation is a translated equivalence of (S).

Sentence ID	Surface forms	Base fomrs	Part of speech	Semantic code	Sentence ID	Surface forms	Base fomrs	Part of speech	Semantic code
(T)	kankou	kankou	NOUN	892	(H2)	kankou (*)	kankou	NOUN	892
[MT Output]	no	no	PARTICLE	-		shi	suru	AUXV	-
	supein	supein	NOUN	709, 719		ni	ni	PARTICLE	-
	ni	ni	PARTICLE	-		supein (*)	supein	NOUN	709, 719
	kuru	kuru	VERB	283, 312		ni	ni	PARTICLE	-
	kappuru	kappuru	NOUN	530		kuru (*)	kuru	VERB	283, 312
	desu	desu	BEVERB	-		kappuru	kappuru	NOUN	530
(H1)	karera	karera	PRONOUN	892		desu	desu	BEVERB	-
	wa	wa	PARTICLE	-	(H3)	kankou (*)	kankou	NOUN	892
	supein (*)	supein	NOUN	709, 719		de	de	PARTICLE	-
	he	he	PARTICLE	-		supein (*)	supein	NOUN	709
	kankou (*)	kankou	NOUN	892		ni	ni	PARTICLE	-
	ni	ni	PARTICLE	-		otozureru	otozureru	VERB	283, 786
	kuru (*)	kuru	VERB	283, 312		hutari(*)	hutari	NOUN	530
	hutari (*)	hutari	NOUN	530		desu	desu	BEVERB	-
	desu	desu	BEVERB	-					

Table 1: The morpheme sequences of both MT output and human translated sentences in Figure 2. (*) indicates that the morpheme can be treated as a keyword defined in Section 2.2.

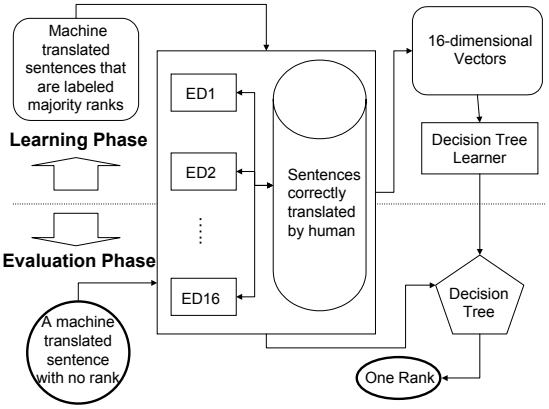

Figure 3: Data flow of our approach

2.2. Edit Distances

All the edit distances, ED1 to ED16, use the same editing operations: insertion, deletion, and replacement /substitution, as in (Wagner, 1974; Thompson 1991; Takezawa et al., 1999; Sugaya et al., 1999; Yasuda et al., 2000; Yasuda et al., 2001). Edit distances, ED9 to ED16 use the additional editing operator: interchange/swap as in (Lowrance, 1975; Su, 1992). The adopted editing cost/weight is the same value, 1, as in (Takezawa et al., 1999; Sugaya et al., 1999; Yasuda et al., 2000; Yasuda et al., 2001).

For ED1, the unit on which the edit operators are applied is a morpheme, while the editing unit in (Lowrance, 1975; Su, 1992) is a character, the editing unit in (Thompson, 1991) is a phrase, and the editing unit in (Takezawa et al., 1999; Sugaya et al., 1999; Yasuda et al., 2000; Yasuda et al., 2001) is a word. For ED1, two morphemes are regarded as being matched if and only if the base form of each morpheme is the same and each part of speech (POS) tag is the same.

	Adaptation of interchange operator	Restriction to content words	Reference to semantic code	Restriction to keywords
ED1(Base)	No	No	No	No
ED2	No	No	No	Yes
ED3	No	No	Yes	No
ED4	No	No	Yes	Yes
ED5	No	Yes	No	No
ED6	No	Yes	No	Yes
ED7	No	Yes	Yes	No
ED8	No	Yes	Yes	Yes
ED9	Yes	No	No	No
ED10	Yes	No	No	Yes
ED11	Yes	No	Yes	No
ED12	Yes	No	Yes	Yes
ED13	Yes	Yes	No	No
ED14	Yes	Yes	No	Yes
ED15	Yes	Yes	Yes	No
ED16	Yes	Yes	Yes	Yes

Table 2: Edit distances

For example, the second column in Table 1 gives simplified morpheme sequences of (T), (H1), (H2), and (H3) in Figure 2. The third and forth columns give, respectively, the base form and the POS tag of the morpheme in the same line. The fifth column for the morphemes correspondent to content words gives their semantic codes. When ED1 is calculated for the pair of (T) and (H1), *kankou* "sightseeing" in (T) and *kankou* "sightseeing" in (H1) are matched because they have the same base form and POS tag. On the other hand, *no* "of" in (T) and *kankou* "sightseeing" in (H1) are not matched because they have the different base forms and POS tags.

The other edit distances ED*i* (*i*=2,…,16) are the extensions of ED1 as follows:

(1) For the edit distances whose third columns are filled with Yes, the only morphemes that are content words are used as the editing units. For

example, in the case of ED5, *kankou* "sightseeing" in (T) is an editing unit while *no* "of" in (T) is ignored as an editing unit.

(2) For the edit distances whose fifth columns are filled with Yes, only morphemes that are content words but are not keyword words are ignored as the editing units. Here, keywords are defined as the content words that appear in a majority number of human-translated sentences. For example, in the case of human-translated sentences: (H1), (H2) and (H3) in Figure 2, *karera* "they" appear only in (H1); therefore *karera* "they" is not a keyword and ED2 does not treat *karera* "they" as an editing unit.

(3) For the edit distances whose fourth columns are filled with Yes, semantic codes of morphemes are used instead of the base forms. For example, in the case of ED3, content words like *kuru* "coming" and *otozureru* "visiting" are compared with semantic codes in stead of the base forms. In the case that some semantic codes are shared like '283' in the semantic codes of both *kuru* "coming" and *otozureru* "visiting", the correspondent morphemes are regard as being matched.

3. Experimental Work

3.1. Experimental Method

The authors evaluated the proposed method on the following two points:

(1) Whether the proposed method is more accurate than the existing methods based on single edit distances, and

(2) How the proposed method can contribute to the evaluation of MT systems.

In order to evaluate the above points, the authors used English-to-Japanese TDMT and bilingual data of English and Japanese, which are specifically transcribed texts (Takezawa, 1999) of real dialogues in the travel arrangement domain. Three hundred forty-three English-to-Japanese-machine-translated sentences were ranked by three native speakers of Japanese. Each of the ranked Japanese sentence consisted of ten words on average. Moreover, the English sentences were translated by twenty-five translators. The resulting total number of correct human-translated sentence was equal to twenty six. Figure 4 shows the number of different human-translated Japanese and its percentage. Let us call different human-translated Japanese "standards".

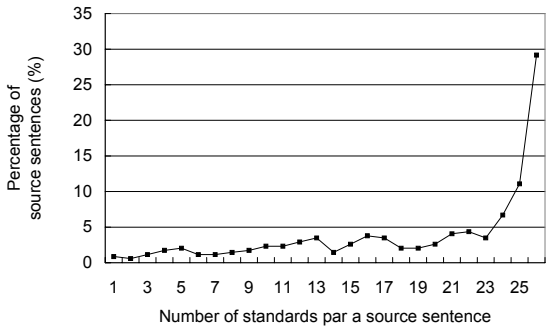

Figure 4: The number of the different human-translated sentences (standards) and its percentage

The tagger used was that of Japanese-to-English TDMT. The semantic codes used were those of Ruigo-Shin-Jiten (Ohno & Hamanishi, 1981). The decision tree learner used was C4.5 (Quinlan, 1993), which is well-known in the machine learning community. The accuracies of automatic ranking method were calculated using the 10-fold cross-validation technique (Mitchell, 1997), which is typically used to evaluate machine learning algorithms.

3.2. Experimental Results

3.2.1. The proposed method V.S. the existing methods

In order to check the first point, the proposed method is compared with discrimination analysis (DA) using either ED1 or ED9 listed in Table 2. ED1 was used in (Thompson, 1991; Takezawa et al., 1999; Sugaya et al., 1999; Yasuda et al., 2000; Yasuda et al., 2001) and ED9 was used in (Su, 1992). Discrimination analysis is one of the typical and well-used classification methods on one-dimensional space.

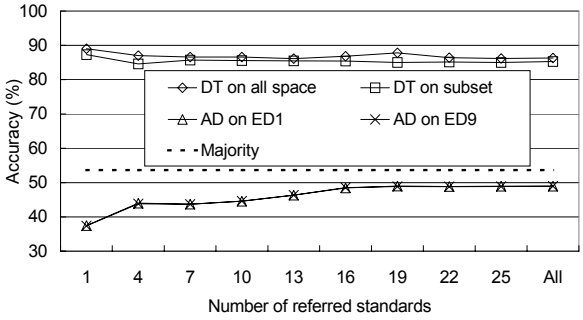

Figure 5: Comparison among the proposed method on all EDs, the proposed methods on partial EDs, and the existing methods, on the closed data

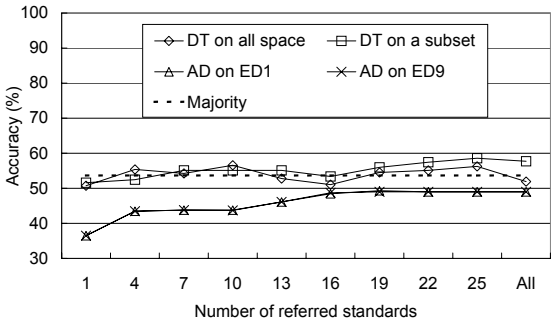

Figure 6: Comparison among the proposed method on all EDs, the proposed methods on partial EDs, and the existing methods, on the open data

Figures 5 and 6 show, respectively, the accuracies on closed data and open data. The horizontal axis shows the number of the referred standards in the edit distance calculation. "DT on all EDs" indicates the results of the proposed method. "DT on partial EDs" indicates the results of the restricted version of the proposed method, which restricted the original sixteen-dimensional vector space to a subspace. That is, "DT on partial EDs" uses only a part of ED*i* (i = 1 … 16) in Table 2. This restriction is called feature selection in the machine learning community. "DA on ED1" indicates the result of DA by using the edit distance ED1. "DA on ED9" indicates the result of DA by using the edit distance ED9. "Majority" indicates the results of always estimating the majority rank: A. This estimation of ranking is hereafter call the default ranking.

Figure 5 shows that both the proposed method "DT on all EDs" and its restricted method "DT on partial EDs" are much more accurate than the single-edit-distance-based ranking methods: "DA on ED1" and "DA on ED9". Moreover, both the proposed method "DT on all EDs" and its restricted method "DT on partial EDs" are more accurate than the default ranking while the single-edit-distance-based ranking methods: "DA on ED1" and" DA on ED9" are less accurate than the simple default ranking.

Figure 6 shows that, while the accuracy of proposed method "DT on all EDs" on open test can be better or worse than the default ranking, its restricted methods "DT on partial EDs" on open test can be almost always better than the default ranking. This means that selection of referred standards is expected to improve the accuracy of the proposed method. And also, the way to identify the most effective combination of edit distances: ED1 to ED16 is expected. These remain open problems.

3.2.2. TDMT quality by manual V.S. TDMT quality by using the proposed method

In order to check the second point, the authors attempted to estimate the TDMT quality by using the proposed method: DT on all EDs. The quality was indirectly calculated from expected confusion matrix of ten confusion matrix in open test.

Figure 7 shows the estimation error, that is,difference between the TDMT quality by the proposed method and TDMT quality by manual. Note that the whole data for the 10-cross-validation was the data with ranks assigned by manual, which was reported in (Sumita, 1999). "A+B" indicates the difference between the sum of the rank A and B that was assigned by the proposed method and the sum of the rank A and B that was assigned by human-evaluators. In the case that the number of the referred standards is equal to 13 or 19, the error is within 3%.

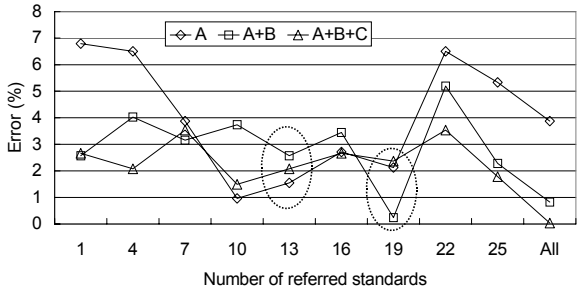

Figure 7: Comparison between EJ-TDMT quality by Manual (Sumita et al., 1999) and EJ-TDMT quality by using the proposed method (DT on all EDs)

Therefore, if the suitable number of referred standards is identified, the proposed method has the potential to estimate the quality of TDMT within several percentage error. In such case, if the automatic ranking method shows that the quality of TDMT is increased much more than 3% by changing something, we can believe that the change is acceptable. On the other hand, the automatic ranking method shows that the quality of TDMT is decreased much more than 3% by changing something, we can believe that the change is not acceptable. In order to claim these points, the authors will extent the experiment by using more large data or by the data in other domains except travel arrangement domain.

4. Conclusion

This paper addressed the problem of automatically ranking output from MT systems. This paper proposed the automatic ranking method that, by using multiple edit distances, encodes machine-translated sentences with a rank assigned by humans into multi-dimensional vectors from which a classifier of ranks is learned as a decision tree. The proposed method assigns a rank to MT output through the learned DT. The proposed method was evaluated using transcribed texts of real conversations in the travel arrangement domain. The proposed method is more accurate than the single-edit-distance-based ranking methods on both closed and open data sets. The proposed method has the potential to estimate the quality of ATR's MT system, TDMT, within several percentage error.

Acknowledgements

The authors thank Kadokawa-Shoten for providing them the Ruigo-Shin-Jiten. They appreciate the members of ATR Spoken Language Translation Research Laboratories for supporting this work. They give special thanks to Mr. Nobutaka Yoshioka and Mr. Hideo Kurihara for cooperating the development of the automatic evaluation system.

References

Furuse, O. et al. (1995). Multi-lingual Spoken-Language Translation Utilizing Translation Examples. In Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium: NLPRS'95 (pp. 544--549).

Furuse, O. & Iida, H. (1996). Incremental Translation Utilizing Constituent Boundary Patterns. In Proceedings of the 16th International Conference on Computational Linguistics: COLING'96 (pp. 412--417). New Brunswick, NJ: The Association for Computational Linguistics (ACL).

King M. (1996). Evaluating Natural Language Processing Systems. Communications of the ACM, 39(1), 73--79.

Mitchell, T.M. (1997). Machine Learning, New York: The McGraw-Hill Companies Inc.

Nagao, M. & Tsujii, J. (1985). Evaluation of J-E Translation Results in Mu-Project (in Japanese). In IPSJ SIG Notes NL47-11 (pp. 79--88). Tokyo: Information Processing Society of Japan (IPSJ).

Lowrance R. & Wagner R.A. (1975). An Extension of the String-to-String Correction Problem. Journal of the ACM, 22(2), 177--183.

Ohno, S. and Hamanishi, M. (1981). Ruigo-Shin-Jiten. Tokyo: Kadokawa shoten Publishing Co LTD.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Su, K. et al. (1992). A New Quantitative Quality Measure for Machine Translation Systems. In Proceedings of the 14th International Conference on Computational Linguistics: COLING'92 (pp. 433--439). New Brunswick, NJ: The Association for Computational Linguistics (ACL).

Sugaya, F. et al. (2000). Evaluation of the ATR-MATRIX Speech Translation Systems with a Pair Comparison Method between the System and Humans. In Proceedings of the 6th International Conference on Spoken Language Processing: ICSLP'00 (pp. 1105--1108).

Sumita, E. et al. (1999). Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach. In Proceedings of MT Summit VII (pp. 229--235). Tokyo, Japan: Asian-Pacific Association for Machine Translation (AAMT).

Takezawa, T. (1999). Building a Bilingual Travel Conversation Database for Speech Translation Research. In Proceedings of the 2nd International Workshop on East-Asian Language Resources and Evaluation - Oriental COCOSDA Workshop'99 (pp. 17--20).

Takezawa, T. et al. (1999). A New Evaluation Method for Speech Translation Systems and a Case Study on ATR-MATRIX from Japanese to English. In Proceedings of MT Summit VII (pp. 299--307). Tokyo, Japan: Asian-Pacific Association for Machine Translation (AAMT).

Thompson, H.S. (1991). Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment. In Proceedings of the Evaluator's Forum (pp. 215--223), Switzerland: ISSCO.

Wagner R.A. (1974). The String-to-String Correction Problem. Journal of the ACM, 21(1), 168--173.

Yamamoto, S. (2000). ATR Spoken Language Translation Laboratories (ATR-SLT) - Towards Speech Translation Technology for the Real World Applications (in Japanese). The Journal of the Acoustical Society of Japan, 56(11), 756--759.

Yasuda, K. et al. (2000). An Automatic Evaluation Method of Translation Capability by DP Matching Using Similar Expressions Queried from a Parallel Corpus (in Japanese). In IEICE Technical Report, NLC2000-63 (pp. 97--102). Tokyo: the Institute of Electronics, Information and communication Engineers (IEICE).

Yasuda, K. et al. (2001). An Automatic Evaluation Method of Translation Quality Using Translation Answer Candidates Queried from a Parallel Corpus. In Proceedings of MT Summit VIII (to appear). Switzerland: The European Association for Machine Translation (EAMT).