Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 1-8.
Association for Computational Linguistics.

Corpus-Centered Computation

Eiichiro SUMITA

ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika, Souraku
Kyoto 619-0288, JAPAN
http://www.atr.co.jp/slt
eiichiro.sumita@atr.co.jp

Abstract

To achieve translation technology that is adequate for speech-to-speech translation (S2S), this paper introduces a new attempt named Corpus-Centered Computation, (abbreviated to C^3 and pronounced *c-cube*). As opposed to conventional approaches adopted by machine translation systems for written language, C^3 places corpora at the center of the technology. For example, translation knowledge is extracted from corpora, translation quality is gauged by referring to corpora and the corpora themselves are normalized by paraphrasing or filtering. High-quality translation has been demonstrated in the domain of travel conversation, and the prospects of this approach are promising due to the benefits of synergistic effects.

1 Introduction

Text-based MT systems are not suitable for speech-to-speech translation (S2S) partly because they have not been designed to cope with the deviations from conventional grammar that characterize spoken language input and partly because they have been designed to be as general as possible to cover as many domains as possible. Consequently, the translation quality is not good[1] enough for S2S purposes. Furthermore, since such systems have been constructed by human experts, the development of machine translation systems and porting them to different domains are expensive and snail-paced processes.

This paper introduces a new attempt named Corpus-Centered Computation, (abbreviated to C^3 and pronounced *c-cube*). C^3 places corpora at the center of the technology, where, for example, translation knowledge is extracted from corpora, translation quality is gauged by referring to corpora, and the corpora themselves are normalized by paraphrasing or filtering.

C^3 has demonstrated its ability to provide high-quality translation. The construction is done by machine, allowing quick and low-cost development.

Section 2 introduces the corpus we are currently dealing with, Section 3 briefly explains our three corpus-based machine translation systems, Section 4 demonstrates the first round of competition between the three systems on the same corpus, Section 5 touches on the automatic selection of the best translation, Section 6 introduces a combination of corpus-based processes, such as translation and paraphrasing, Section 7 discusses the implications of our approach, and finally Section 8 concludes the paper.

2 The Corpus

We are aiming at the development of a S2S system to be used in place of a phrasebook by foreign tourists. **Table 1** shows the English part of some sample translation pairs from our Japanese and English corpus.

Table 2 compares our corpus with two other spoken language corpora developed by ATR (Takezawa, T. et al., 2002) and Verbmobil (Ney, H. et al., 2000).[2] Our corpus has the shortest

[1] For our travel domain, a famous translation system on the WEB between Japanese and English produced a good translation for only about 10~20% of our test sentences.

[2] J, E, and G stand for Japanese, English, and German.

average sentence length. On the other hand, it is rich in topics and thus has the largest vocabulary and volumes.

Table 1. Sample English sentences in our corpus

I want to buy a roll of film.
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.

Table 2. Comparison with other bilingual spoken language corpora

	ATR/ dialogue	Our corpus	Verbmobil
#Sent.	16,110	204,108	58,332
#Word	(J) 231,267 (E) 181,263	(J) 1,689,449 (E) 1,235,747	(G) 519,523 (E) 549,921
Voc.	(J) 4,895 (E) 4,032	(J) 19,640 (E) 15,374	(G) 7,940 (E) 4,673
Length	(J) 14.3 (E) 11.3	(J) 8.3 (E) 6.1	(G) 8.9 (E) 9.4

3 Three Corpus-based Machine Translation Systems

Research on corpus-based translation is a growing trend and has become indispensable to the MT industry.

There are two main strategies used in corpus-based translation:

1. Example-Based Machine Translation (EBMT): EBMT uses the corpus directly. EBMT retrieves the translation examples that are best matched to an input expression and adjusts the examples to obtain the translation (Nagao, 1984; Somers, 1999).
2. Statistical Machine Translation (SMT): SMT learns models for translation from corpora and dictionaries and searches at run-time for the best translation according to the models (Brown et al., 1993; Knight, 1997; Ney et al., 2000).

We have developed two EBMT systems and one SMT system.

3.1 An EBMT, D^3

Sumita (2001) proposed D^3 (Dp-match Driven transDucer). The characteristics of D^3 are different from previous EBMT approaches: a) Most EBMT proposals assume syntactic parsing and bilingual tree-banks, but D^3 does not; b) Most EBMT proposals divide the translation process in two, i.e. learning of translation patterns in advance and application of the translation patterns, but D^3 generates translation patterns on the fly according to the input and the retrieved translation examples as needed.

As shown in **Figure 1**, our language resources are [i] a bilingual corpus, in which sentences are aligned beforehand; [ii] a bilingual dictionary, which is used for generating target sentences; and [iii] thesauri of both languages, which are used for incorporating the semantic distance between words into the distance between word sequences. Furthermore, [ii] and [iii] are also used for word alignment.

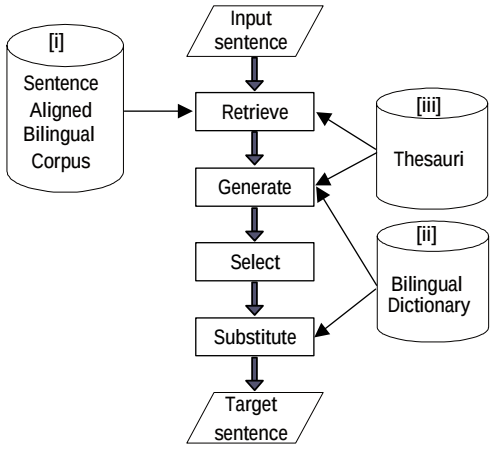

Figure 1. Configuration

Suppose we are translating a Japanese sentence into English. Let's review the process with a simple sample below. The Japanese input (1-j) is translated into English (1-e) by utilizing (2-e), whose source (2-j) is similar to (1-j). The common parts are unchanged, and the different portions are substituted by consulting a bilingual dictionary.

;;;A Japanese input
(1-j) iro/ga/ki/ni/iri/masen

;;; Japanese part of an example in corpus [i]
(2-j) dezain/ga/ki/ni/iri/masen
;;; English part of an example in corpus [i]
(2-e) I do not like the design.

;;; the English output
(1-e) I do not like the color.

We retrieve the most similar source sentence of examples from a bilingual corpus. For this, we use DP-matching, which tells us the distance between word sequences, *dist* while giving us the matched portions between the input and the example. According to equation [1], *dist*, is calculated. The counts of the Insertion (I), Deletion (D), and substitution operations are summed. Then, this total is normalized by the sum of the lengths of the source and example sequences. According to equation [2], substitution considers the semantic distance, *SEMDIST*, between two substituted words.

$$[1]\ dist = \frac{I + D + 2\sum SEMDIST}{L_{input} + L_{example}}$$

$$[2]\ SEMDIST = \frac{K}{N}$$

Figure 2 illustrates the SEMDIST calculation between "potato" and "beef." The *most specific common abstraction* of the two words is "ingredients," as shown in boldface. The SEMDIST is *K* divided by *N* in the figure. *K* is the level of the most specific common abstraction of two words, and *N* is the height of the thesaurus.

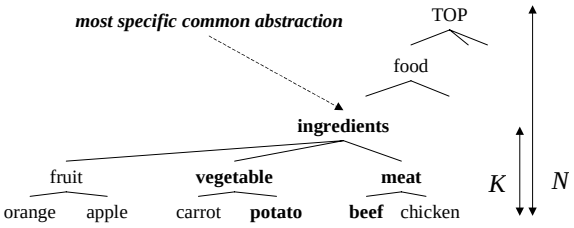

Figure 2. SEMDIST calculated using thesaurus

Let us show the latest performance of D[3]. The translation speed is sufficiently fast in that the average translation time is 0.04 seconds per sentence with the 200 K corpus shown in Section 2. The translation quality is so high that the method can achieve a TOEIC score[3] of 750. This is equivalent to the average score of a Japanese businessperson in overseas department of Japanese corporations.

In brief, D[3] uses DP-matching, which features the semantic distance between words. D[3] has demonstrated good quality and short turnaround in travel conversations such as those in a phrasebook.

3.2 EBMT and SMT based on Hierarchical Phrase Alignment (HPA)

Here we intoroduce Hierarchical Phrase Alignment (HPA) and its application to EBMT and SMT.

3.2.1 Hierarchical Phrase Alignment (HPA)

This subsection introduces a new phrase alignment approach (Imamura, 2001) called Hierarchical Phrase Alignment (HPA).

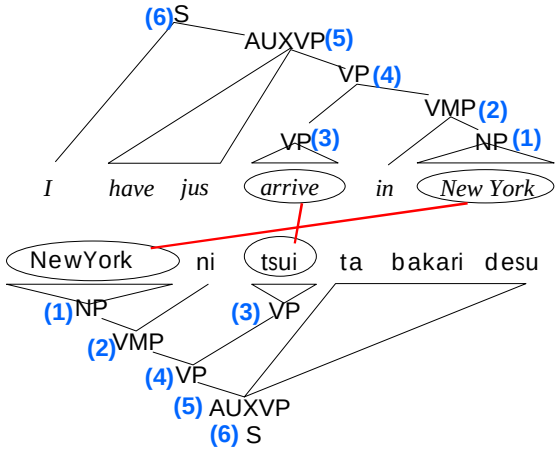

Figure 3. Bilingual trees and alignment

First, two sentences are tagged and parsed independently. This operation obtains two syntactic trees. Next, words are linked by the word alignment program. Then, HPA retrieves *equivalent phrases* that satisfy two conditions: 1) words in the pair correspond with no deficiency and no excess; 2) the phrases are of the same syntactic category.

Let's look at a sample pair of a Japanese

[3] The TOEIC (Test of English for International Communcation) test is an English language proficiency test for people whose native language is not English (http://www.chauncey.com/). The Total score ranges from 10 to 990. ATR has developed a method to measure the TOEIC score of a machine translation system. (Sugaya et al., 2000)

tree and the corresponding English tree (**Figure 3**). The retrieval of equivalent phrases is done in a bottom-up fashion. First, the syntactic node pair that consists of only the 'New York' and 'NewYork' link, having the same syntactic category, is retrieved. Then, NP(1) and VMP(2) are found. Next, the syntactic node pair that consists of only the 'arrived' and 'tsui' link, having the same syntactic category, is retrieved. Then, VP(3) is found. Finally, the syntactic node pairs that include two word links having the same syntactic category are retrieved. Then VP(4), AUXVP(5), and S(6) are found. Accordingly, six *equivalent phrases* are hierarchically extracted.

3.2.2 EBMT based on HPA

Imamura (2002) proposed an application of HPA in EBMT called HPA-based translation (HPAT). HPAed bilingual trees include all information necessary to automatically generate transfer patterns. Translation is done according to transfer patterns using the TDMT engine (Sumita et al., 1999), our previous EBMT system. First, the source part of transfer patterns are utilized, and source structure is obtained. Second, structural changes are performed by mapping source patterns to target patterns. Finally, lexical items are inserted by referring to a bilingual dictionary, and then a conventional generation is performed. HPAT achieved about 70% accuracy.

3.2.3 SMT based on HPA

Statistical machine translation (SMT) represents a translation process as a noisy channel model that consists of a source-channel model and a language model of the target language.

The translation model is based on word-for-word translation and limited to allow only one channel source word to be aligned from a channel target word. Although phrasal correspondence is implicitly implemented in some translation models by means of distortion, careful parameter training is required.

In addition, the training procedure relies on the EM algorithm, which can converge to an optimal solution but does not assure the global maximum parameter assignment. Furthermore, the numbers of parameters represent the translation models, so that easily suffered from the over-fitting problem. In order to overcome these problems, simpler models, such as word-for-word translation models (Brown et al., 1993) or HMM models (Och et al., 2000), have been introduced to determine the initial parameters and to bootstrap the training.

We have proposed two methods to overcome the above problems by using HPA. (1) The first method converts the hierarchically aligned phrasal texts into a pair of sequences of chunks of words, treating the word-for-word translation model as a chunk-for-chunk translation model. (2) The second method computes the parameters for the translation model from the computed phrase alignments and uses the parameters as a starting point for training iterations.

The experimental results on Japanese-to-English translation indicated that the model trained from the parameters derived from the HPA could improve the quality of translation (Watanabe et al., 2002a).

4 Competition between the Three MTs on Same Corpus

4.1 Competition Conditions

We used the corpus shown in Section 2, which is a collection of Japanese sentences and their English translations, typically found in phrasebooks for foreign tourists. We lemmatized and POS-tagged both the Japanese and English sentences. A quality evaluation was done for the test set consisting of 510 sentences selected randomly from the above corpus, and the remaining sentences were used for learning and verification.

We also used a bilingual dictionary previously developed for TDMT. The size of the dictionary is 24,658 words. We used thesauri whose hierarchies are based on the Kadokawa Ruigo-shin-jiten (Ohno and Hamanishi, 1984). The size of the Japanese thesaurus is 21,608 and that of the English thesaurus is 11,359.

4.2 Results

SMT has been applied to language pairs of similar European languages. We implemented SMT for translation between Japanese and English, which are dissimilar in many points such as word order. **Table 3** shows the accuracy of our SMT system. The four ranks are defined as follows (Sumita et al., 1999): (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. It worked in both J-to-E and E-to-J directions [4] in spite of the negative opinions previously expressed.

Table 3. SMT worked for J and E

Rank(s)	*A*	*A+B*	*A+B+C*
SMT(JE)	25%	46%	64%
SMT(EJ)	41%	48%	57%

We implemented two EBMT systems, D^3 and HPAT, using the same corpus. D^3 and HPAT surpassed SMT in the travel conversation task (**Tables 3 and 4**).

Table 4. EBMTs on the same corpus

Rank(s)	*A*	*A+B*	*A+B+C*
D^3(JE)	47%	66%	77%
HPAT(EJ)	50%	61%	71%

Finally, it became clear that word-based SMT, a revival of the direct method of the '50s, is suitable for pairs of European languages but not for Japanese and English. This is because word-based SMT cannot capture the major differences such as word order between Japanese and English.

Several organizations (Yamada et al., 2001; Alshawi et al., 2000) are pursuing syntax-based SMT. We plan to join the race. *Which is suitable for Japanese and English, syntax-based SMT or EBMT?*

5 Combination of Evaluation and Translation

We are researching automatic evaluation of machine translation outputs and multiple paradigms for machine translation simultaneously. Together, they have synegistic effects as explained below.

5.1 Automatic Quality Evaluation Using Corpus

Translation quality has conventionally been evaluated by hand. Likewise, we have evaluated the outputs of our translation systems subjectively with four ranks from 'good' to 'bad': A, B, C, and D (Sumita et al., 1999).

Such subjective evaluation by ranking, however, is taxing on both time and resources. If automatic evaluation methods are inexpensive, fast, and sufficiently accurate, then such automatic evaluation methods would prove beneficial.

Conventional approaches to automatic evaluation include methods (Su, 1992; Yasuda et al., 2001) that automatically assign one of several ranks to MT output according to a *single* edit distance between an MT output and a correct translation example.

To improve performance, we proposed 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 decision tree (Akiba et al., 2001).

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 accurately estimate the quality of outputs of machine translation systems.

5.2 Multiple-engine Machine Translation System

Every researcher has his own way of acquiring translation knowledge by generalizing translation instances in a corpus. Our approach is no exception to this rule. Our MTs are based on different paradigms, different development styles, and different development periods. This results in various behaviors for each input sentence, and the translation rank of a given input sentence changes system-by-system.

Table 5. Sample of transaltion variety with quality rank

o-shiharai wa genkin desu ka kurejitto kaado desu ka
[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?

[4] For this test set, the accuarcy of SMT is at least twice as good as that of a famous conventional machine translation system on the WEB.

We show a sample of different English translations obtained by three systems for a Japanese sentence (**Table 5**). The brackets show the quality rank judged by a human translator.

Translation systems gain A-ranked translations in different subsets of input sentences as illustrated in **Figure 4**. Thus, we could obtain a large increase in accuracy by using an "ideal" MT, if it were possible to select the best one of the three different translations for each input sentence.

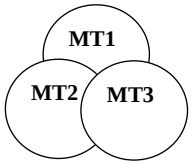

Figure 4. Subsets of input sentences whose translation is A-ranked

We are investigating methods to utilize techniques of automatic evaluation for *selector* (**Figure 5**).

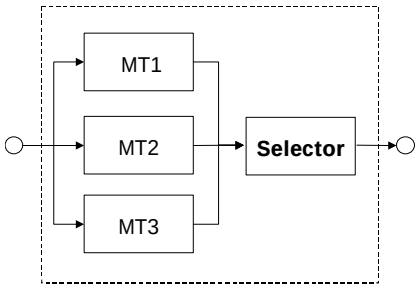

Figure 5. Selector for multi-engine MT

In our pilot experiment, our selectors (Akiba et al., 2002; Yasuda et al., 2002) outperformed not only the component systems but also a conventional selector using N-gram (Callison-Burch et al., 2001).

6 Combination of Paraphrasing and Translation

We are automating extraction of paraphrase knowledge from a bilingual corpus. In this section, we introduce its application to improve the performance of corpus-based translation by using SMT as a touchstone.

6.1 Extraction of Synonymous Expressions

We propose an automatic paraphrasing method that exploits knowledge from bilingual corpora (Shimohata et al., 2002).

Synonymous expressions are defined as a sequence of variant words with surrounding common words. The expressions are extracted from bilingual corpora by the following procedures (**Figure 6**):

1. Collect sentences that share the same translation in another language. The accumulated sentences are defined as *synonymous sentences*.
2. For all pairs of *synonymous sentences*, apply DP-matching and collect sequences of words, *synonymous expressions that* consist of variant words preceded/followed by common words.
3. Remove pairs of *synonymous expressions* with a frequency lower than a given threshold.
4. Cluster the pairs of *synonymous expressions* by transitive relation.

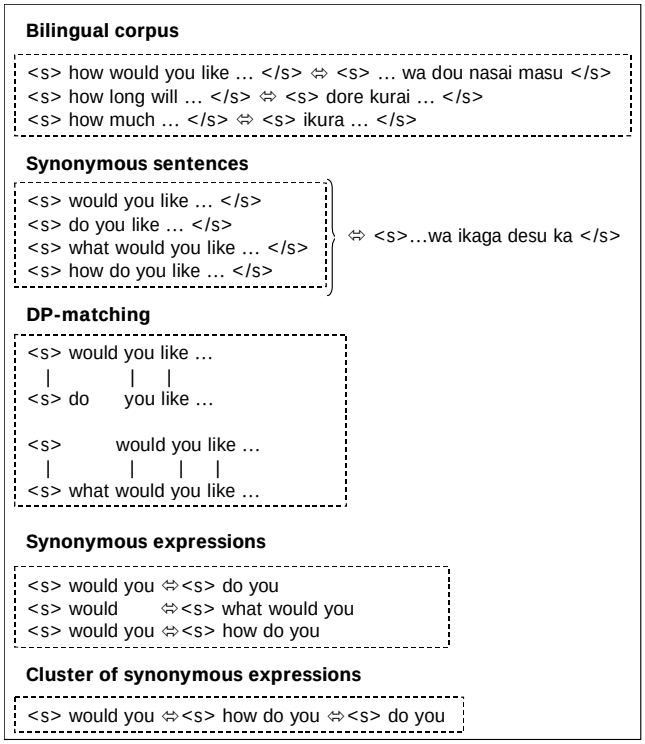

Figure 6. Extraction of synonymous expressions

6.2 Corpus Normalization

After the acquisition of clusters of synonymous expressions, normalization is carried out by transforming the expressions into major ones,

selected according to their frequency in the corpora. For instance, the cluster obtained consists of the expressions '<s> would you,' '<s> how do you' and '<s> do you.' Suppose that an expression '<s> do you' occurred most frequently in a given corpus, an input 'how do you like your coffee' could be normalized into 'do you like your coffee.'

6.3 SMT on Normalized Corpus

Statistical approach to machine translation demands huge bilingual corpora in good quality and broad coverage. However, such an ideal corpus is not usually available: one may contain a sufficiently large number of samples, for instance, derived from web pages with translations, but these may not be well-aligned or have low translation quality. Others may consist of high-quality translations but have a limited number of examples. In addition, the variety of possible translations makes it even harder to estimate parameters for statistical-based machine translation.

We propose a way to overcome these problems by creating a corpus that consists of normalized expressions, expressions with less variety, through automated paraphrasing (Watanabe et al., 2002b). As described above, by the method of transforming *target* sentences of a given bilingual corpus into a *normalized form* is expected to improve the parameter estimation for a statistical machine translation model. The normalization method proposed above locally replaces word sequences, hence will not affect the syntactical coherence. Therefore, normalization will not affect the distortion model, which accounts for reordering of bilingual texts. In addition, reduction of the vocabulary size will greatly help improve the parameter estimation for lexical models.

The experimental results on Japanese-to-English translation indicated that the SMT created on the target normalized sentences reduced word-error-rate from 66% to 58%.

7 Discussions

7.1 Forecasting from the Obtained Performance

As a component of C^3, D^3 has achieved a high TOEIC score. We foresee much higher scores for C^3 because it features a multi-engine and selector scheme, which is an easy, quick and low-cost method of improving total performance, since there is no need to investigate the messy relationships between resources and processes of the component systems by hand.

7.2 Backcasting from the Future S2S System in the Real World

We are aiming to develop technologies for S2S that are usable in real-world environments. No one knows what the real world will be, but there is no doubt that an S2S system should deal with variations in length and expressions beyond our corpus that explained in Section 2. In other words, we divided our "real-world" goal into three sub-goals; (1) translation of short and edited sentences; (2) translation of long sentences; (3) translation of short but non-edited sentences; and (4) combining solutions for these sub-goals seamlessly.

Since we are centering our approaches on corpora, we are developing corpora for achieving sub-goals (1), (2) and (3) as reported in (Takezawa et al., 2002; Sugaya et al., 2002) .

For sub-goal (1), we are using a selector for multiple engines, for sub-goal (2), we have to devise methods to chunk long sentences into appropriate translation units, and for sub-goal (3) we need a powerful automatic paraphraser.

8 Conclusions

Our attempt called C^3 places corpora at the center of S2S technology. All components of C^3 are corpus-based as shown in the paper. If we have sufficient volumes of sentence-aligned bilingual corpora, we would be able to build a high-quality MT. Corpus-based processes for such tasks as translation, evaluation, and paraphrasing have synergistic effects. Therefore, we are optimistic about the progress of components and their integration in C^3.

Acknowledgements

The author's heartfelt thanks go to Kadokawa-Shoten for providing the Ruigo-Shin-Jiten. The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

References

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

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 (to appear)

Alshawi, H., Bangalore, S. and Douglas, S. 2000 Learning Dependency Translation Models as Collections of Finite-State Head Transducers, Computational Linguistics, 26 (1), pp. 45--60.

Brown, P., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Laffetry, J. D., Mercer, R. L. and Roossin, P. S. 1993 A Statistical Approach to Machine Translation, Computational Linguistics 16, pp. 79--85

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

Imamura, K. 2001 Hierarchical phrase alignment harmonized with parsing, Proc. of NLPRS, pp. 377--384

Imamura, K. 2002 Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment, Proc. of TMI

Knight, K. 1997, Automating Knowledge Acquisition for Machine Translation, AI Magazine, 18 (4), pp. 81--96

Nagao, M. 1984 A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in A. Elithorn and R. Banerji (eds), Artificial and Human Intelligence, Amsterdam: North-Holland, pp. 173--180.

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

Och et al. 2000. Improved statistical alignment models. In Proc. of ACL, pp. 440--447, Hong Kong, China, October

Ohno, S. and Hamanishi, M. 1984. Ruigo-Shin-Jiten, Kadokawa, Tokyo (in Japanese)

Shimohata, M. and Sumita, E. 2002 Automatic paraphrasing based on parallel corpus for normalization, Proc. of LREC

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

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

Sugaya, F., Takezawa, T., Yokoo, A., Sagisaka, Y, and Yamamoto, S. 2000 Evaluation of the ATR-MATRIX Speech Translation System with a Pair Comparison Method Between the System and Humans, Proc. of ICSLP, pp. 1105--1108

Sugaya, F. et al. 2002 Proposal for a very large-corpus acquisition method by registering in tree-structure form, Proc. of LREC

Sumita, E. 2001 Example-based machine translation using DP-matching between word sequences, Proc. of DDMT (ACL), pp. 1--8

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, pp. 229--235

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

Yamada et al. 2001 A Syntax-Based Statistical Translation Model. Proc. of ACL, France

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

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

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

Watanabe, T., Shimohata, M. and Sumita, E. 2002b Statistical Machine Translation Based On Paraphrased Corpora, Proc. of LREC