Automatic Construction of Machine Translation Knowledge Using Translation Literalness

Kenji Imamura, Eiichiro Sumita
ATR Spoken Language Translation Research Laboratories
Seika-cho, Soraku-gun, Kyoto, Japan
{kenji.imamura,eiichiro.sumita}@atr.co.jp

Yuji Matsumoto
Nara Institute of Science and Technology
Ikoma-shi, Nara, Japan
matsu@is.aist-nara.ac.jp

Abstract

When machine translation (MT) knowledge is automatically constructed from bilingual corpora, redundant rules are acquired due to translation variety. These rules increase ambiguity or cause incorrect MT results. To overcome this problem, we constrain the sentences used for knowledge extraction to "the appropriate bilingual sentences for the MT." In this paper, we propose a method using translation literalness to select appropriate sentences or phrases. The translation correspondence rate (TCR) is defined as the literalness measure.

Based on the TCR, two automatic construction methods are tested. One is to filter the corpus before rule acquisition. The other is to split the acquisition process into two phases, where a bilingual sentence is divided into literal parts and the other parts before different generalizations are applied. The effects are evaluated by the MT quality, and about 4.9% of MT results were improved by the latter method.

1 Introduction

Along with the efforts made to accumulate bilingual corpora for many language pairs, quite a few machine translation (MT) systems that automatically construct their knowledge from corpora have been proposed (Brown et al., 1993; Menezes and Richardson, 2001; Imamura, 2002). However, if we use corpora without any restriction, redundant rules are acquired due to translation varieties. Such rules increase ambiguity and may cause inappropriate MT results.

Translation variety increases with corpus size. For instance, large corpora usually contain multiple translations of the same source sentences. Moreover, peculiar translations that depend on context or situation proliferate in large corpora. Our targets are corpora that contain over one hundred thousand sentences.

To reduce the influence of translation variety, we attempt to control the bilingual sentences that are appropriate for machine translation (here called "controlled translation"). Among the measures that can be used for controlled translation, we focus on translation literalness in this paper. By restricting bilingual sentences during MT knowledge construction, the MT quality will be improved.

The remainder of this paper is organized as follows. Section 2 describes the problems caused by translation varieties. Section 3 discusses the kinds of translations that are appropriate for MTs. Section 4 introduces the concept of translation literalness and how to measure it. Section 5 describes construction methods using literalness, and Section 6 evaluates the construction methods.

2 Problems Caused by Translation Variety

First, we describe the problems inherent in bilingual corpora when we automatically construct MT knowledge.

2.1 Context/Situation-dependent Translation

Some bilingual sentences in corpora depend on the context or situation, and these are not always correct in different contexts.

For instance, the English determiner '*the*' is not generally translated into Japanese. However, when a human translator cannot semantically identify the following noun, a determinant modifier such as '*watashi-no* (my)' or '*sono* (its)' is supplied.

As an example of a situation-dependent translation, the Japanese sentence "*Shashin wo tot-te itadake masu ka?* (Could you take our photograph?)" is sometimes translated into an English sentence as "*Could you press this shutter button?*" This translation is correct from the viewpoint of meaning, but it can only be applied when we want a photograph to be taken. Such examples show that most context/situation-dependent translations are non-literal.

MT knowledge constructed from context/situation-dependent translations cause incorrect target sentences, which may contain omissions or redundant words, when it is applied to an inappropriate context or situation.

2.2 Multiple Translations

Generally speaking, a single source expression can be translated into multiple target expressions. Therefore, a corpus contains multiple translations even though they are translated from the same source sentence. For example, the Japanese sentence "*Kono toraberaazu chekku wo genkin ni shite kudasai*" can be translated into English any of the following sentences.

- *I'd like to cash these traveler's checks.*
- *Could you change these traveler's checks into cash?*
- *Please cash these traveler's checks.*

These translations are all correct. Actually, the corpus of Takezawa et al. (2002) contains ten different translations of this source sentence. When we construct MT knowledge from corpora that contain such variety, redundant rules are acquired. For instance, a pattern-based MT system described in Imamura (2002) acquires different transfer rules from each multiple translations, although only one rule is necessary for translating a sentence. Redundant rules increase ambiguity or decrease translation speed (Meyers et al., 2000).

3 Appropriate Translation for MTs

3.1 Controlled Translation

Controlled language (Mitamura et al., 1991; Mitamura and Nyberg, 1995) is proposed for monolingual processing in order to reduce variety. This method allows monolingual texts within a restricted vocabulary and a restricted grammar. Texts written by the controlled language method have fewer semantic and syntactic ambiguities when they are read by a human or analyzed by a computer.

A similar idea can be applied to bilingual corpora. Namely, the expressions in bilingual corpora should be restricted, and "translations that are appropriate for the MT" should be used in knowledge construction. This approach assumes that context/situation-dependent translations should be removed before construction so that ambiguities in MT can be decreased. Restricted bilingual sentences are called controlled translations in this paper.

The following measures are assumed to be available for controlled translation. First three measures are for each of the bilingual sentences in the corpus and the fourth measure is for the whole corpus:

- **Literalness:** Few omissions or redundant words appear between the source and target sentences. In other words, most words in the source sentence correspond to some words in the target sentence.
- **Context-freeness:** Source word sequences correspond to the target word sequences independent of the contextual information. With this measure, partial translation can be reused in other sentences.
- **Word-order Agreement:** The word order of a source sentence agrees substantially with that of a target sentence. This measure ensures that the cost of word order adjustment is small.
- **Word Translation Stability:** A source word is better translated into the same target word through the corpus.

  For example, the Japanese adjectival verb '*hitsuyoo-da*' can be translated into the En-

glish adjective '*necessary*,' the verb '*need*,' or the verb '*require*.' It is better for an MT system to always translate this word into '*necessary*,' if possible.

Effective measures of controlled translation depend on MT methods. For example, word-level statistical MT (Brown et al., 1993) translates a source sentence with a combination of word transfer and word order adjustment. Thus, word-order agreement is an important measure. On the other hand, this is not important for transfer-based MTs because the word order can be significantly changed through syntactic transfer. A transfer-based MT method using the phrase structure is studied here.

3.2 Base MT System

We use Hierarchical Phrase Alignment-based Translator (HPAT) (Imamura, 2002) as the target transfer-based MT system. HPAT is an new version of Transfer Driven Machine Translator (TDMT) (Furuse and Iida, 1994). Transfer rules of HPAT are automatically acquired from a parallel corpus, but those of TDMT were constructed manually.

The procedure of HPAT is briefly described as follows (Figure 1). First, phrasal correspondences are hierarchically extracted from a parallel corpus using Hierarchical Phrase Alignment (Imamura, 2001). Next, the hierarchical correspondences are transferred into patterns, and transfer rules are generated. At the time of translation, the input sentence is parsed by using source patterns in the transfer rules. The MT result is generated by mapping the source patterns to the target patterns. Ambiguities, which occur during parsing or mapping, are solved by selecting the patterns that minimize the semantic distance between the input sentence and the source examples (real examples in the training corpus).

3.3 Appropriate Translation for Transfer-based MT

In order to verify effective measures of controlled translation for transfer-based MTs, we review the fundamentals of TDMT in this section.

TDMT was trained by human rule writers. They selected bilingual sentences from a corpus one by

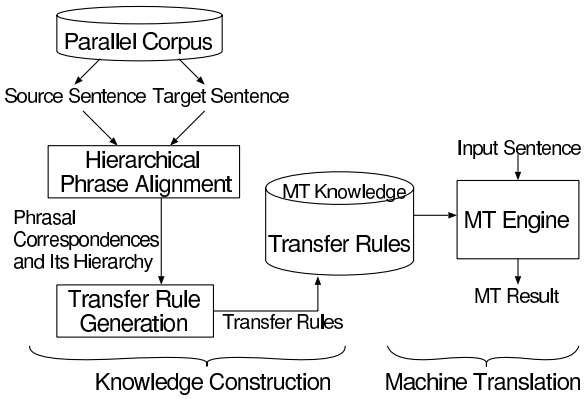


Figure 1: Overview of HPAT: Knowledge Construction and Translation Process

one and added or arranged the transfer rules in order to translate the sentences. The target sentences were then rewritten with the aim of minimizing the number of transfer rules. We believe that this way of rewritten translation is appropriate examples for TDMT.

We compared 6,304 bilingual sentences rewritten for an English-to-Japanese version of TDMT and the original translations in the corpus [1]. The statistics in Table 1 show that the following measures are effective for transfer-based MT. Note that these data were calculated from the results of morphological analysis and word alignment (c.f., Section 6). The correspondences output from the word aligner are called word links.

Literalness Focusing on the number of linked target words, the value of the rewritten translations is considerably higher than that of the original translations. This result shows that the words of source sentences are translated into target words more directly in the case of the rewritten translations. Thus, the rewritten translations are more literal.

Word Translation Stability Focusing on the number of different words in the target language and the mean number of translation words, both values of the rewritten translations are lower than those of the original translations. This is because

[1] When TDMT translates input sentences already trained, the MT results become identical to the objective translations for the rule writer. Therefore, the rewritten translations were acquired by translating trained sentences by TDMT.

	Rewritten Translations	Original Translations
# of Linked Target Words	28,300 words (49.5%)	20,722 words (34.0%)
# of Different Words in Target Language	3,107 words	3,601 words
Mean # of Translation Words per Source Word	1.51 trans./word	1.94 trans./word
Mean Context-freeness (# of Word Link = 4)	4.45	4.21

Table 1: Comparison of TDMT Training Translations and Original Translations

the rule writers rewrote translations to make target words as simple as possible, and thus the variety of target words was decreased. In other words, the rewritten translations are more stable from the viewpoint of word translation.

Context-freeness Mean context-freeness in Table 1 denotes the mean number of word-link combinations in which word sequences of the source and the target contain word links only between their constituents (cross-links are allowed). If a bilingual sentence can be divided into many translation parts, this value become high. This value depends on the number of word links. When it is calculated only from the sentences that contain four word links, the value of the rewritten translations is higher than that of the original translations.

4 Translation Literalness

We particularly focus on the literalness among the controlled translation measures in order to reduce the incorrect rules that result from context/situation-dependent translations. Word translation stability and context freeness must serve as countermeasures for multiple translations, since they ensure that word translations and structures are steady throughout the corpus. However, the reduction of incorrect translations is done prior to the reduction of ambiguities.

4.1 Literalness Measure

A literal translation means that source words are translated one by one to target words. Therefore, a bilingual sentence that has many word correspondences is literal. The word correspondences can be acquired by referring to translation dictionaries or using statistical word aligners (e.g., (Melamed, 2000)).

However, not all source words always have an exact corresponding target word. For example, in the case of English and Japanese, some prepositions are not translated into Japanese. On the contrary, the preposition '*after*' may be translated into Japanese as the noun '*ato*.' These examples show that some functional words have to be translated while others do not. Thus, literalness is not determined only by counting word correspondences but also by estimating how many words in the source and target sentences have to be translated.

Based on the above discussion, the translation literalness of a bilingual sentence is measured by the following procedure. Note that a translation dictionary is utilized in this procedure. The dictionary is automatically constructed by gathering the results of word alignment at this time, though hand-made dictionaries may also be utilized. In this process, we assume that one source word corresponds to one target word.

1. Look up words in the translation dictionary by the source word. T_s denotes the number of source words found in the dictionary entries.

2. Look up words in the dictionary by target words. T_t denotes the number of target words found in the definition parts of the dictionary.

3. If there is an entry that includes both the source and target word, the word pair is regarded as the word link. L denotes the number of word links.

4. Calculate the literalness with the following equation, which we call the Translation Correspondence Rate (TCR) in this paper.

$$TCR = \frac{2L}{T_s + T_t} \quad (1)$$

The TCR denotes the portion of the directly translated words among the words that should be translated. This definition is bi-directional,

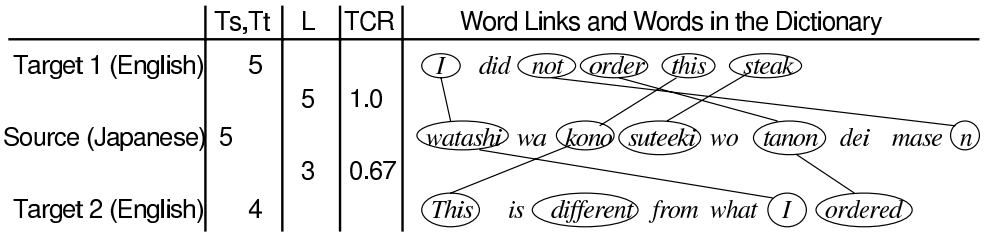


Figure 2: Example of Measuring Literalness Using Translation Correspondence Rate
(Circled words denote words found in the dictionary.
Lines between sentences denote word links.)

so omission and redundancy can be measured equally. Moreover, the influence of the dictionary size is low because the words that do not appear in the dictionary are ignored.

For example, suppose that a Japanese source sentence (Source) and its English translations (Targets 1 and 2) are given as shown in Figure 2. Target 1 is a literal translation, and Target 2 is a non-literal translation, while the meaning is equivalent. When the circled words are those found in the dictionary, T_s is five, and T_t of Target 1 is also five. There are five word links between Source and Target 1, so the TCR is 1.0 by Equation (1).

On the other hand, in the case of Target 2, four words are found in the dictionary ($T_t = 4$), and there are three word links. Thus, the TCR is $\frac{2*3}{5+4} \simeq 0.67$, and Target 1 is judged as more literal than Target 2.

The literalness based on the TCR is judged from a tagged result and a translation dictionary. In other words, 'deep analyses' such as parsing are not necessary.

5 Knowledge Construction Using Translation Literalness

In this section, two approaches for constructing translation knowledge are introduced. One is bilingual corpus filtering, which selects highly literal bilingual sentences from the corpus. Filtering is done as preprocessing before rule acquisition. The other is split construction, which divides a bilingual sentence into literal and non-literal parts and applies different generalization strategies to these parts.

5.1 Bilingual Corpus Filtering

We consider two approaches to corpus filtering.

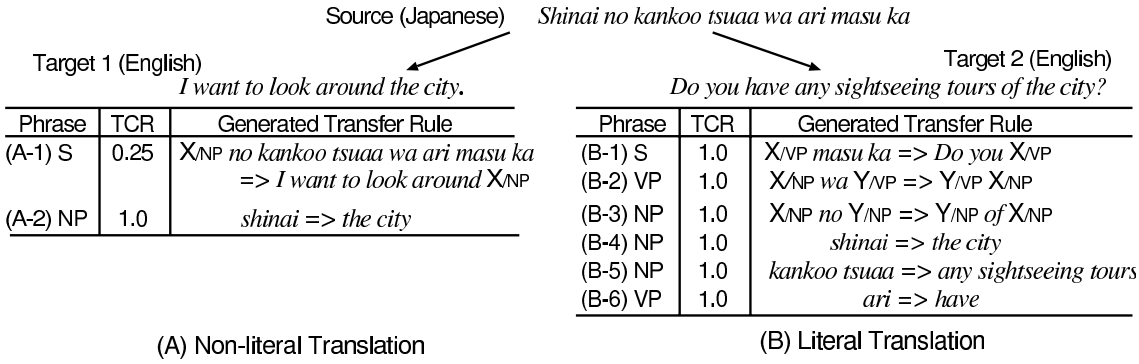
Filtering Based on Threshold A partial corpus is created by selecting bilingual sentences with TCR values higher than a threshold, and MT knowledge is constructed from the extracted corpus. By making the threshold higher, the coverage of MT knowledge will decrease because the size of the extracted corpus becomes smaller.

Filtering Based on Group Maximum First, sentences that have the identical source sentence are grouped together, and a partial corpus is created by selecting the bilingual sentences that have the maximal TCR from each group. As opposed to filtering based on a threshold, all source sentences are used for knowledge construction, so the coverage of MT knowledge can be maintained. However, some context/situation-dependent translations remain in the extracted corpus when only one non-literal translation is in the corpus.

5.2 Split Construction into Literal and Other Parts

The TCR can be calculated not only for sentences but also for phrases. In the case of filtering, the coverage of the MT knowledge is decreased by limiting translation to highly literal sentences. However, even though they are non-literal, such sentences may contain literal translations at the phrase level. Thus, the coverage can be maintained if we extract literal phrases from non-literal sentences and construct knowledge from them.

A problem with this approach is that non-literal bilingual sentences sometimes contain idiomatic



Phrase	TCR	Generated Transfer Rule
(A-1) S	0.25	X/NP *no kankoo tsuaa wa ari masu ka* => *I want to look around* X/NP
(A-2) NP	1.0	*shinai* => *the city*

(A) Non-literal Translation

Phrase	TCR	Generated Transfer Rule
(B-1) S	1.0	X/VP *masu ka* => *Do you* X/VP
(B-2) VP	1.0	X/NP *wa* Y/VP => Y/VP X/NP
(B-3) NP	1.0	X/NP *no* Y/NP => Y/NP *of* X/NP
(B-4) NP	1.0	*shinai* => *the city*
(B-5) NP	1.0	*kankoo tsuaa* => *any sightseeing tours*
(B-6) VP	1.0	*ari* => *have*

(B) Literal Translation

Figure 3: Examples of Generated Rules for Japanese-to-English Translation (A) from Non-literal Translation by Split Construction (B) from Literal Translation.

translations that should not be translated literally. For example, the Japanese greeting "*Hajime mashi te*" should be translated into "*How do you do*," not into its literal translation, "*For the first time*." Such idioms are usually represented by a long word sequence.

To cope with literal and idiomatic translations, a sentence is divided into literal and non-literal parts, and a different construction is applied. Short rules, which are more generalized and easier to reuse, are generated from the literal parts. Long rules, which are more strict in their use in MT, are generated from the non-literal parts. The procedure is described as follows.

1. Phrasal correspondences are acquired by Hierarchical Phrase Alignment.
2. The hierarchy is traced from top to bottom, and the literalness of each correspondence is measured. If the TCR is equal to or higher than the threshold, the phrase is judged as a literal phrase and the tracing stops before reaching the bottom.
3. If the phrase is literal, transfer rules that include its lower hierarchy are generalized.
4. If the top structure (i.e., entire sentence) is not literal, a rule is generated in which only the literal parts are generalized.

For example, suppose that different target sentences from the same source are given as shown in Figure 3. The phrase (A-1)S has low TCR, but the TCR of the noun phrase pair '*shinai*' and '*the city*' has 1.0. Thus, the phrase (A-2)NP is generalized, and the long transfer rule (A-1)S is generated from the non-literal translation. On the contrary, the TCR of the top phrase (B-1)S is 1.0, so all phrases in (B) are generalized and totally six rules are generated. The rules generated from literal translations are general, and they will be used for the translation of the other sentences.

Thus, by using the split construction, rules like templates are generated from non-literal translations and primary rules for transfer-based MT are generated only from literal phrases. Rules generated from non-literal translations are used only when the input word sequence exactly matches the sequence in the rule. In other words, they are hardly used in different contexts.

6 Translation Experiments

In order to evaluate the effect of literalness in MT knowledge construction, we constructed knowledge by using the methods described in Section 5 and evaluated the MT quality of the resulting English-to-Japanese translation.

6.1 Experimental Settings

Bilingual Corpus We used as the training set 149,882 bilingual sentences from the Basic Travel Expression Corpus (Takezawa et al., 2002). This corpus is a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists. There are many bilingual sentences in

which the source sentences are the same but the targets are not. About 13% of different English sentences have multiple Japanese translations.

Translation Dictionary: Extraction of Word Correspondence For word correspondences that occur more than nine times in the corpus, statistical word alignment was carried out by a similar method to Melamed (2000). When words for which the correspondence could not be found remain, a thesaurus (Ohno and Hamanishi, 1984) was used to create correspondences to the words of the same group. A translation dictionary was constructed as a collection of the word correspondences. The accuracy of this word aligner is about 90% for precision and 73% for recall by a closed test of content words.

Evaluation for MT Quality We used the following two methods to evaluate MT quality.

1. **Automatic Evaluation**

   We used BLUE (Papineni et al., 2002) with 10,150 sentences that were reserved for the test set. The number of references was one for each sentence, and a range from uni-gram to four-gram was used.

2. **Subjective Evaluation**

   From the above-mentioned test set, 510 sentences were evaluated by paired comparison. In detail, the source sentences were translated using the base rule set created from the entire corpus, and the same sources were translated using the rules constructed with literalness. One by one, a Japanese native speaker judged which MT result was better or that they were of the same quality. Subjective quality is represented by the following equation, where I denotes the number of improved sentences and D denotes the number of degraded sentences.

$$\text{Subj. Quality} = \frac{I - D}{\text{\# of test sentences}} \quad (2)$$

6.2 MT Quality vs. Construction Methods

The level of MT quality achieved by each of the construction methods is compared in Table 2. Coverage of exact rules denotes the portion of sentences that were translated by using only the rules that require the source example to exactly match the input sentence. In addition, the threshold $TCR \geq 0.4$ was used for filtering because it was experimentally shown to be the best value. In the case of split construction, we used the extracted corpus after filtering based on the group maximum, and phrases that were $TCR \geq 0.8$ were judged to be literal phrases.

First, focusing on the filtering, the subjective qualities or the BLEU scores are better than the base in both methods. Comparing the threshold with the group maximum, the BLEU score is increased by the group maximum. The coverage of the exact rules is higher even if the corpus size decreases. Filtering based on the group maximum improves the quality while maintaining the coverage of the knowledge.

Although we used a high-density corpus where many English sentences have multiple Japanese translations, the quality improved by only about 1%. It is difficult to significantly improve the quality by bilingual corpus filtering because it is difficult to both remove insufficiently literal translations and maintain coverage of MT knowledge.

On the other hand, the BLEU score and the subjective quality both improved in the case of split construction, even though the coverage of the exact rules decreased. In particular, the subjective quality improved by about 4.9%. Incorrect translations were suppressed because the rules generated from non-literals are restricted when the MT system applies them.

In summary, all construction methods helped to improve the BLEU scores or the subjective qualities; therefore, construction with translation literalness is an effective way to improve MT quality.

7 Conclusions

In this paper, we proposed restricting the translation variety in bilingual corpora by controlled translation, which limits bilingual sentences to the appropriate translations for MT. We focused on literalness from among the various measures for controlled translation and defined a Translation Correspondence Rate for calculating literalness.

Less literal translations could be removed by fil-

	Entire Corpus Base	Filtering		Split Construction ($TCR \geq 0.8$)
		Threshold ($TCR \geq 0.4$)	Group Maximum	
# of Translations (Size Ratio)	149,882 (100%)	129,069 (86.1%)	118,686 (79.2%)	118,686 (79.2%)
Coverage of Exact Rules	67.1 %	65.1 %	66.3 %	60.6 %
BLEU Score	0.225	0.224	0.231	**0.240**
Subjective Quality		+1.4 %	+0.6 %	**+4.9 %**
# of Improved Sentences		26	31	116
# of Same Quality (Same Results)		465 (421)	451 (393)	303 (182)
# of Degraded Sentences		19	28	91

Table 2: MT Quality vs. Construction Methods

tering according to the TCR, and this slightly improved the MT quality.

The TCR is capable of measuring literalness not only for bilingual sentences but also for phrases. In other words, a bilingual sentence can be divided into literal phrases and other phrases. Using this feature, sentences were divided into literal parts and non-literal parts, and transfer rules that could be applied with strong conditions were generated from the non-literal parts. As a result, MT quality as judged by subjective evaluation improved in about 4.9% of the sentences.

Word translation stability and context-freeness were also effective measures. MT quality is expected to be further improved by using these measures because they reduce multiple translations.

Acknowledgment

The research reported here is 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

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Osamu Furuse and Hitoshi Iida. 1994. Constituent boundary parsing for example-based machine translation. In *Proceedings of COLING-94*, pages 105–111.

Kenji Imamura. 2001. Hierarchical phrase alignment harmonized with parsing. In *Proceedings of NLPRS-2001*, pages 377–384.

Kenji Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *TMI-2002*, pages 74–84.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, June.

Arul Menezes and Stephen D. Richardson. 2001. A best first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the 'Workshop on Example-Based Machine Translation' in MT Summit VIII*, pages 35–42.

Adam Meyers, Michiko Kosaka, and Ralph Grishman. 2000. Chart-based translation rule application in machine translation. In *Proceedings of COLING-2000*, pages 537–543.

Teruko Mitamura and Eric H. Nyberg. 1995. Controlled English for knowledge-based MT: Experience with the KANT system. In *Proceedings of TMI-95*.

Teruko Mitamura, Eric H. Nyberg, and Jamie G. Carbonell. 1991. An efficient interlingua translation system for multi-lingual document production. In *Proceedings of MT Summit III*, pages 55–61.

Susumu Ohno and Masato Hamanishi. 1984. *Ruigo-Shin-Jiten*. Kadokawa, Tokyo (in Japanese).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL-2002*, pages 311–318.

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. In *LREC 2002*, pages 147–152.