自 然 言 語 処 理 82－5
（1991 3. 15）

用例主導型機械翻訳

隅田　英一郎・飯田　仁

ATR自動翻訳電話研究所

概要

本稿では用例主導型機械翻訳方式について述べる。本方式では、入力に類似した用例（原文と訳文の対）を、用例データベースから検索し、それに基づいて翻訳する。本方式は以下の長所を持つ。(1) 改良容易性、(2) 信頼度付与、(3) 高速、(4) ロバスト性、(5) 翻訳者の技能の有効利用。本稿では、特に、日本語の「の」の英語への翻訳に、本方式を適用した実験結果を報告し、また、本方式の広範な適用可能性について議論する。

Experiments and Prospects of Example-Based Machine Translation

Eiichiro SUMITA and Hitoshi IIDA

ATR Interpreting Telephony Research Laboratories
Sanpeidani, Inuidani, Seika-cho, Souraku-gun, Kyoto 619-02, JAPAN

Abstract

EBMT (Example-Based Machine Translation) is proposed. EBMT retrieves similar examples (pairs of texts and their translations), adapting the examples to translate a new input. EBMT has the following features: (1) is easily upgraded; (2) produces a reliability factor; (3) is quick; (4) is robust; and (5) well utilizes translator expertise. An experiment translating Japanese noun phrases of the form "N_1 no N_2" to English is described. Broad applicability of EBMT is discussed using text corpus.

1 Introduction

Machine Translation requires handcrafted and complicated knowledge[1]. Conventional machine translation systems **use rules as the knowledge.** This framework is called Rule-Based Machine Translation (RBMT). It is difficult to scale up from a toy program to a practical system because it is difficult to build such a large-scale rule-base. It is also difficult to improve translation performance because the effect of adding a new rule is difficult to anticipate, and because translation using a large-scale rule-based system is time-consuming.

In order to conquer these problems in machine translation, **the use of a database of examples (pairs of source texts and their translations) as the knowledge** has been instituted[2, 3, 4, 5a, 5b]. The translation mechanism retrieves similar examples from the database, adapting the examples to translate the new source text. This framework is called Example-Based Machine Translation (EBMT).

In order to advocate that EBMT is promising and can be a real breakthrough in MT technology, this paper makes use of ATR's linguistic database of spoken Japanese with English translations, whose corpus is conversation about registering for an international conference[6].

This paper is composed as follows: (1) It explains in detail our pilot system which translates Japanese noun phrases of the form "N_1 no N_2" to English noun phrases. An average success ratio of about 78% has been achieved in the experiments. Moreover, it can be made better as discussed below; (2) it analyzes how broadly EBMT is applicable by using actual data.

The rest of this section examines **RBMT** and **EBMT** from various aspects and sets pointers for the following sections.

(1) Improvement

In RBMT, it is too difficult to keep all rules consistent because they are mutually dependent. RBMT is not easily upgraded.

However, EBMT has no rules, and the use of an example is relatively localized. Improvement is effected simply by inputting appropriate examples to the database. EBMT is easily upgraded. The experiment in section 4.3.2 has shown this property of EBMT: **the more examples we have, the better the quality.**

(2) Reliability Factor

One of the main reasons users dislike RBMT systems is the so-called "poisoned cookie" problem. RBMT has no device to compute the reliability of the result. In other words, users of RBMT cannot trust any RBMT translation, because it can be wrong without any such indication from the system.

In EBMT, a reliability factor is assigned to the translation result according to the distance between the input and found similar examples[see the experiment in section 4.3.3]. In addition to this, retrieved examples which are similar to the input have considerable persuasive power in convincing users that the translation is accurate.

(3) Translation speed

RBMT translates slowly in general because RBMT is really a large-scale rule-based system, which consists of analysis, transfer, and generation modules using syntactic rules, semantic restrictions, structural transfer rules, word selections, generation rules, and so on. For example, the Mu system has about 2,000 rewriting rules and word selection rules for about 70,000 lexical items[7].

As recently pointed out[8], conventional RBMT systems have been biased toward syntactic, semantic, and contextual analysis which consumes considerable computing time. However, such deep analysis is not always necessary or useful for translation.

In contrast with this, deep semantic analysis is avoided in EBMT because it is assumed that **translations can be done without deep understanding.** EBMT directly returns a translation without reasoning through a long chain of rules [see sections 2 and 4].

There is fear that retrieval from a large-scale example database is very slow. However, it can be accelerated effectively by both indexing[9] and parallel computing. Their effects are independent and are multiplied. Consequently, EBMT is acceptabely efficient.

(4) Robustness

RBMT works on exact-match reasoning. RBMT fails to translate when it has no knowledge that matches the input exactly.

EBMT works on best-match reasoning. EBMT intrinsically translates in a fail-safe way[see sections 2 and 4].

(5) linguistics + translator expertise

Formulating linguistic rules for RBMT is a difficult job and needs a linguistically trained staff. Moreover, linguistics does not deal with all phenomena occurring in real text[10].

However, examples necessary for EBMT are easy to obtain because a large number of texts, and their translations as a realization of translator expertise, dealing with all real phenomena, are available. Moreover, as electronic publishing increases, more and more texts will be machine-readable[17]. Please see further discussion in section 3.

Section 2 explains the basic idea of EBMT, section 3 explains how broadly EBMT be applicable, sections 4 and 5 give a rationale for section 3, i.e., section 4 illustrates an experiment on translating noun phrases of the form "N_1 no N_2" in detail, and section 5 studies other phenomena through actual data from our corpus.

2 Basic Idea of EBMT

2.1 Basic Flow

In this section, the basic idea of EBMT, which is general and applicable to many phenomena dealt with by machine translation, is shown.

• Example Database(portion for "kireru"[cut / be sharp])
(1) houtyou wa kireru -> The kitchen knife cuts.
(2) kanojyo wa kireru -> She is sharp.
• Input
katyou wa kireru -> ?
• Retrieval of similar examples
(Syntax) Input = (1), (2)
(Semantics) katyou /≈ houtyou, katyou ≈ kanojyo
(Total) Input ≈(2)
• Output -> The chief is sharp.

Figure 1 Mimicking similar examples

Figure 1 shows the basic flow of EBMT using translation of "kireru"[cut / be sharp]. From here on, the literal English translations are bracketed.

Examples (1) and (2) are Japanese to English examples in the database.

Retrieval of examples similar to the Japanese input sentence is performed. Syntactically, the input is similar to Japanese sentences (1) and (2). However, semantically, "katyou" [chief] is far from "houtyou" [kitchen knife]. But, "katyou" [chief] is similar to "kanojyo" [she]. In other words, the input is similar to example sentence (2).

By translating the input mimicking a similar example (2), we finally get "The chief is sharp".

2.2 Distance

Retrieval of similar examples can be rendered by **selection of examples whose distance is minimum.** To define the best distance is the essence of EBMT and is not yet solved completely. However, one promising definition will be shown in section 4.2.1.

From similar examples retrieved, EBMT generates the most likely translation with a reliability factor based on distance and frequency. If there is no similar example within the given threshold, EBMT tells the user that it cannot translate the input.

3 Broad Applicability

EBMT is applicable to well-known or new linguistic phenomena which are regarded as difficult in conventional RBMT.

When one of the following conditions is present, EBMT is suggested.

(a) It is difficult to formulate translation rules.
(b) The general rule is not meaningful because it is special.
(c) Ordinal compositionality from words does not hold[2, 17, 18].

This is a list (not exhaustive) of phenomena in J-E translation, which are suitable for EBMT:

• optional cases ("~ de", "~ ni",...)
• subordinate conjunction ("~ ba ~", "~ nagara ~", "~ tara ~",...,"~ baai ~",...)
• noun phrase of the form "N_1 no N_2"
• sentence of the form "N_1 wa N_2 da"
• sentence lacking main verb (ex. sentence of the form "~ o-negaishimasu")
• fragmental expression ("hai", "sou-desu", "wakarimashita",...)[8]
• modality represented by sentence end ("~tai-no-desu-ga", "~se-te-itadaki-masu", ...)[8]
• simple sentence[3]

This paper gives a detailed experiment for "N_1 no N_2" in section 4 and prospects for other phenomena, "N_1 wa N_2 da" and "~ o-negaishimasu" in section 5.

It is not yet clear whether EBMT can/should deal with the whole process of translation. We assume that there are many kinds of phenomena: some are suitable for EBMT; Others are not. In other words, they are suitable for RBMT.

We have decided to prove the usefulness of EBMT by **integrating EBMT with RBMT.** It would be more acceptable for users, if RBMT is first introduced as a base system which can translate totally, then **incrementally improve** its translation performance by attaching EBMT components. We proposed a practical method of integration in previous papers[5a, 5b].

4 EBMT for "N_1 no N_2"

4.1 The problem

"N_1 no N_2" is a frequent Japanese noun phrase form. "no" in the "N_1 no N_2" is a Japanese adnominal particle. There are other vairiants, including "deno", "karano", madeno" and so on.

Roughly speaking, Japanese noun phrases of the form "N_1 no N_2" correspond to English noun phrases of the form "N_2 of N_1" as shown in the upper examples in Figure 2.

Japanese	*English*
youka **no** gogo	the afternoon **of** the 8th
kaigi **no** mokuteki	the object **of** the conf.
kaigi **no** sankaryou	the application fee **for** the conf.
	?the application fee **of** the conf.
kyouto **deno** kaigi	the conf. **in** Kyoto
	?the conf. **of** Kyoto
issyukan **no** kyuuka	a week'**s** holiday
	?the holiday **of** a week
mittsu **no** hoteru	three hotels
	?hotels **of** three

Fugure 2 Variations in translation of "N_1 no N_2"

However, "N_2 of N_1" does not always provide a natural translation as shown in the lower examples in Figure 2. Some translations are too broad in meaning to interpret, others are almost ungrammatical. For example, the fourth one, "the conference of Kyoto", could be misconstrued as "the conference about Kyoto", and the last one, "hotels of three", is not English. Natural translations often require prepositions other than "of", or no preposition at all. In only about one-fifth of "N_1 no N_2" occurrences in our domain would "N_2 of N_1" be the most appropriate English translation. We cannot use any particular preposition as an effective default value.

No rules for selecting the most appropriate translation for "N_1 no N_2" have yet been found. In other words, the condition (a) in section 3 holds. Selecting the translation for "N_1 no N_2" is still an important and difficult problem in J-E translation.

In contrast with the proceding research analyzing "N_1 no N_2"[11,12], deep semantic analysis is avoided because it is assumed that **translations can be done without deep understanding.** This assumption supports EBMT which directly returns a translation by adapting the examples without reasoning through a long chain of rules.

4.2 Implementation

The EBMT system consists of two databases: an example database and a thesaurus, and three translation modules: analysis, example-based transfer, and generation(Figure 3).

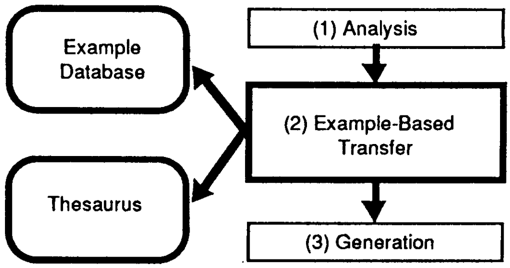

Figure 3 System Configuration

Examples are extracted from ATR's linguistic database of spoken Japanese with English Translations. The corpus is conversation about registering for an international conference[6].

The thesaurus is used to calculate the semantic distance between the content words in the input and those in the examples. The hierarchy of the thesaurus is in accordance with the thesaurus of everyday Japanese written by Ohno and Hamanishi[13].

Figure 4 illustrates the translation procedure with an actual sample. First, morphological analysis is performed for the input "kyouto[Kyoto] deno kaigi [conference]". In this case, syntactical analysis is not necessary. Second, similar examples are retrieved from the database. The top 5 similar examples are shown. Note that the top 3 examples have the same distance and that they are all translated with "in". Third, using this rationale, EBMT generates "the conference in Kyoto".

Analysis

kyouto **deno** kaigi

Example-Based Transfer

d	*Japanese*	*English*
0.4	toukyou **deno** taizai	the stay **in** Tokyo
0.4	honkon **deno** taizai	the stay **in** Hongkong
0.4	toukyou **deno** go-taizai	the stay **in** Tokyo
1.0	oosaka no kaigi	the conf. in Osaka
1.0	toukyou no kaigi	the conf. in Tokyo

Generation

the conf. **in** Kyoto

Figure 4 Translation Procedure

4.2.1 Distance Calculation

The distance used when retrieving examples is essential and is explained here in detail.

Here we suppose the input and examples in the database are represented in the same data structure, the list of the attribute's values. We refer to them and their i-th attribute as I, E and I_i, E_i, respectively.

The attributes of the current target, "N_1 no N_2", are as follows: for nouns, "N_1" and "N_2", the lexical subcategory of the noun, the existence of a prefix or suffix, and its semantic code in the thesaurus; for the adnominal particle "no", the kinds of variants, "deno", "karano", "madeno" and so on. Here, for simplicity, only the semantic code and the kind of adnominal are considered.

Distances are calculated using the following two expressions[5a, 5b]:

$$(1)\ d(I,E)=\sum_i d(I_i,E_i) * w_i$$

$$(2)\ w_i=\sqrt{\sum_{t.p.} (\text{freq. of t. p. when } E_i=I_i)^2}$$

The attribute distance, $d(I_i, E_i)$, and the weight of attribute, w_i, are explained in the following sections.

(a) Attribute Distance

For the attribute of the adnominal particle "no", the distance is 0 or 1 depending on whether or not they match exactly, for example, d("deno","deno") = 0 and d("deno", "no") = 1.

For semantic attributes, however, the match is partial and the distance varies between 0 and 1. Semantic distance $d(0 \leq d \leq 1)$ is determined by the Most Specific Common Abstraction(MSCA)[14] from the thesaurus abstraction hierarchy. For example, when the thesaurus is (n+1) layered, (k/n) is assigned to the concepts in the k-th layer from the bottom. For example, as shown with the broken line in Figure 5, the MSCA("kaigi" [conference], "taizai" [stay]) is "koudou" [actions] and the distance is 2/3.

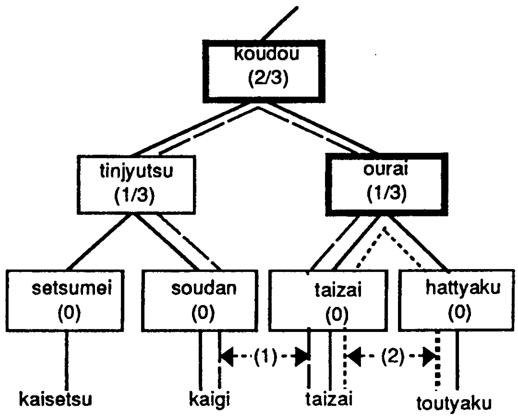

Figure 5 Thesaurus(portion)

(b) Weight of Attribute

The weight of the attribute is the degree to which the attribute influences the selection of the translation pattern(t.p.). We adopt the expression (2) used by Stanfill for memory-based reasoning[15], to implement the intuition.

In Figure 6, all the examples whose E_2 = "deno" are translated with the same preposition, "in". This implies that when E_2= "deno", E_2 is an attribute which heavily influences the selection of the translation pattern. In contrast to this, the translation patterns of examples whose E_1="timei"[place], are varied. This implies that when E_1 = "timei[place]", E_1 is an attribute which is less influential on the selection of the translation pattern.

t.p.	*freq.*	*t.p.*	*freq.*	*t.p.*	*freq.*
B in A	12/27	B in A	3/3	B	9/24
AB	4/27			AB	9/24
B from A	2/27			B in A	2/24
BA	2/27			A's B	1/24
...	...			...	...
B to A	1/27			B on A	1/24
(E1=timei)		*(E2=deno)*		*(E3=soudan)*	

Figure 6 Weight of the i-th attribute

According to the expression (2) , weights for attributes, E_1 and E_2 are as follows:

$w_1 = \sqrt{((12/27)^2+(4/27)^2+...+(1/27)^2)} \approx 0.49$

$w_2 = \sqrt{((3/3)^2} = 1.0$

4.3 Experiment

The current number of words in the corpus is about 300,000 and the number of examples is 2,550. The collection of examples from another domain is in progress.

4.3.1 Jackknife Test

In order to roughly estimate the translation performance, we conducted a jackknife experiment.

We partitioned the example database(2,550) in groups of one hundred, then used one set as input(100) and translated them with the rest as an example database (2,450) and repeated this 25 times.

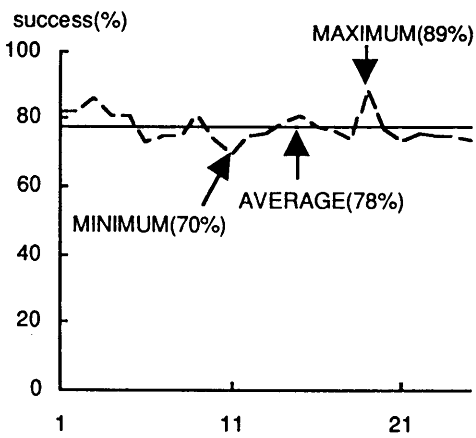

Figure 7 Result of jackknife test

Figure 7 shows that **the average success rate is 78%, the minimum 70% and the maximum 89%**[see section 4.3.4].

It is difficult to fairly compare this result with a translation by the existing MT system. However, it is considered that a conventional system can at best output the major translation pattern, for example, "B of A", as the default. In that case, the success ratio may be about 20% .

4.3.2 Relationship between success and number of examples

Figure 8 shows the relationship between the success rate and the number of examples. There are three cases: maximum, average, and minimum in the previous jackknife test. This graph shows that, in general, **the more examples we have, the better the quality**[see section 4.3.4].

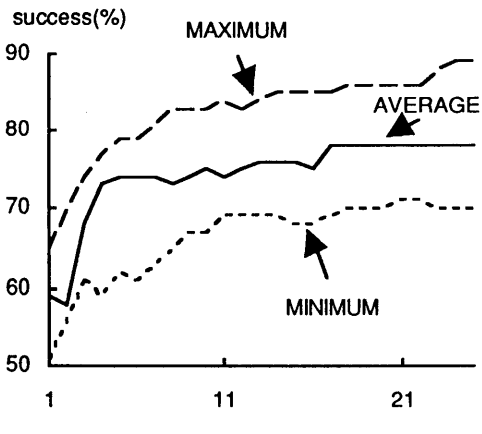

Figure 8 Success and no. of examples

4.3.3 Relationship between success and distance

Figure 9 shows the relationship between the success rate and the distance between the input and the most similar examples retrieved.

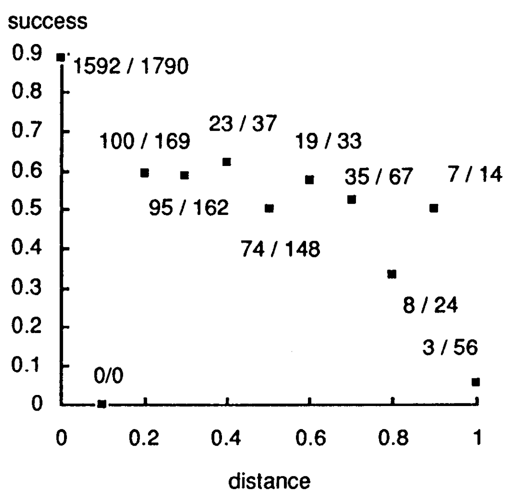

Figure 9 Success and distance

This graph shows that, **in general, the smaller the**

distance, the better the quality. In other words, EBMT provides the distance between the input and the retrieved examples as a reliability factor.

4.3.4 A look at successes and failures

Here, typical successes are shown: (1) the noun phrase "kyouto-eki [Kyoto-station] no o-mise [store]" is translated according to the translation pattern "B at A" while the similar noun phrase, "kyouto[Kyoto] no shiten [branch]" is translated according to the translation pattern "B in A"; (2) the noun phrase of the form "N_1 no hou" is frequent in spoken Japanese. It is translated according to the translation pattern "A", in other words, the second noun is omitted.

We are now investigating the result carefully and striving to improve the success rate.

(a) About half of the failures are caused by a lack of similar examples. They are easily solved by adding appropriate examples.

(b) The rest are caused by the existence of similar examples: (1) equivalent but different examples are retrieved, for example, examples of the form, "B at A" and "B of A" for "waga-sya [our company] no yaku'in[executive]". They can be regarded as successful or the distance calculation may be changed to handle this problem; (2) The current distance calculation is not the best and dissimilar examples are retrieved. (1) is one of the main reasons the graphs (Figures 7 and 8) are up-and-down.

5 Phenomena other than "N_1 no N_2"

This section studies the phenomena, "N_1 wa N_2 da" and "~ o-negaishimasu" with the same corpus used in the previous section.

5.1 "N_1 wa N_2 da"

The sentence of the form "N_1 wa N_2 da" is called a "da" sentence. Here "N_1" and "N_2" are nouns, "wa" is a topical particle, and "da" is a kind of verb which, roughly speaking, is the English copula "be".

The correspondences between "da" sentences and English are exemplified in Figure 10. Mainly, "N_1 wa N_2 da" corresponds to "N_1 be N_2" like (a-1) ~ (a-4).

However, sentences like (b) ~ (e) cannot be translated according to the translation pattern "N_1 be N_2".

We explain the typical example (d) here. There is no Japanese counterpart of "payment should be made by". The English sentence has explicit modality, passive, verb make, and its object, payment, while the Japanese sentence has no such correspondences. This translation cannot be made in a compositional way, depending on the target words which are selected from a normal dictionary. It is difficult to formulate rules for the translation or to explain how the translation is made. The conditions (a) and (c) in section 3 hold. Conventional approaches lead to understanding "da" sentences deeply by using contextual and extra-linguistic information. However, we have many translations at hand which are the result of human translators' understanding. Translation can be made by mimicking similar examples.

(a) N_1 be N_2	
watshi[I]	jyonson[Johnson]
kotira[this]	jimukyoku[secretariat]
denwa-bango[tel-no.]	06-951-0866[06-951-0866]
sanka-hi[fee]	85,000-en[85,000 yen]
(b) N_1 cost N_2	
yokousyuu[proc.]	30,000-en[30,000 yen]
(c) for N_1, the fee is N_2	
kigyou[companies]	85,000-en[85,000 yen]
(d) payment should be made by N_2	
hiyou[fee]	ginnkou-furikomi[bank-transfer]
(e) the conference will end on N_2	
saisyuu-bi[final day]	10gatsu12niti[12th Oct.]

Figure 10 Examples of "N_1 wa N_2 da"

Moreover, example (e) is special, i.e., idiomatic. The condition (c) in section 3 holds.

The distribution of N_1 and N_2 in the examples of our corpus varies for each case. Study 2-tuples of nouns, (N_1, N_2). N_2s of (a-4), (b) and (c) are similar, i.e., *prices*. However N_1s are not similar to each other. N_1s of (a-4) and (d) are similar, i.e., *fee*. However the N_2s are not similar to each other. Thus, EBMT is applicable.

5.2 "~ o-negaishimasu"

Figure 11 exemplifies the correspondences between a sentence of the form "~ o-negaishimasu" and English.

(a) may I speak to N	jimukyoku[secretariat] **o**
(b) please give me N	go-jyuusyo[address] **o**
(c) please pay by N	genkin[cash] **de**
(d) yes, please	hai
(e) thank you	yoroshiku

Figure 11 Examples of "~ o-negaishimasu"

Translations in examples (b) and (c) are rendered by finding the missing elements, *give me* and *pay by* , respectively. The conditions (a) and (c) in section 3 hold. Usually, this kind of supplement is done by contextual analysis. However, the connection between the missing elements and the noun in the examples are strong enough to reuse, because they are the product of a combination of translator expertise and domain specific restriction.

Examples (a), (d) and (e) are idiomatic expressions. The condition (c) holds.

In the same way as the "da" sentence, the distribution of the noun and the particle in the examples of our corpus varies for each case. EBMT is applicable.

6 Concluding Remarks

EBMT (Example-Based Machine Translation) has been proposed. EBMT retrieves similar examples (pairs of source texts and their translations), adapting the examples to translate a new source text.

EBMT has the following features: (1) is easily upgraded; (2) produces a reliability factor; (3) is quick; (4) is robust; and (5) well utilizes translator expertise.

The feasibility of EBMT has been shown by implementing a system which translates Japanese noun phrases of the form "N_1 no N_2" to English noun phrases. The result of the experiment was encouraging. Broad applicability of EBMT is discussed with data from the text corpus. The system has been written in Common Lisp, and is running on a Genera 7.2 at ATR.

References

1. Nirenburg, S., Machine Translation, Cambridge University Press, (1987), 350.
2. Nagao, M. , A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in Artificial and Human Intelligence, ed. A. Elithorn and R. Banerji, North-Holland, (1984), 173-180.
3. Sato, S. and Nagao, M. , Memory-Based Translation , reprint of WGNL 70-9, IPSJ, (1989)(in Japanese).
4. Sadler, V. , Translating with a Simulated Bilingual Knowledge Bank(BKB), BSO/Research, (1989).

5a. Sumita, E., Iida, H. and Kohyama, H., Translating with Examples: A New Approach to Machine Translation, Proc. of the third international conference on theoretical and methodological issues in machine translation of natural languages, Texas, (1990), 203-212.

5b. Sumita, E. Iida, H. and Kohyama, H. , Example-based Approach in Machine Translation, InfoJapan '90, (1990), Part 2: 65-72.

6. Ogura, K., Hashimoto, K. , and Morimoto, T. , Object-Oriented User Interface for Linguistic Database, Proc. of Working Conference on Data and Knowledge Base Integration, University of Keele, England, (1989).
7. Nagao, M. ,Tsujii, J. , Nakamura, J. , Machine Translation from Japanese into English, Proceedings of the IEEE, 74, 7, (1986).
8. Furuse, O., Sumita, E.and Iida, H., A method for realizing Transfer-Driven Machine Translation, reprint of WGNL 80-8, IPSJ, (1990)(in Japanese).
9. Sumita, E. and Tsutsumi, Y. , A translation aid system using flexible text retrieval based on syntax-matching, Proc. of the second international conference on theoretical and methodological issues in machine translation of natural languages, CMU, Pittsburgh, (1988).
10. Nagao, M.(chair), Language engineering : The real bottle neck of natural language processing, Proc. of Coling '88, (1988).
11. Shimazu, A. , Naito, S. , and Nomura, H. , Semantic structure analysis of Japanese noun phrases with adnominal particles, Proc. of 25th ACL, (1987), 123-130.
12. Hirai, M. and Kitahashi, T. , A Semantic classification of noun modifications in Japanese sentences and their analysis, reprint of WGNL 58-1, IPSJ, (1986)(in Japanese).
13. Ohno, S. and Hamanishi, M. , Ruigo-Shin-Jiten, Kadokawa, (1984), 932(in Japanese).
14. Kolodner, J. and Riesbeck, C. , Case-Based Reasoning, tutorial textbook of 11th IJCAI, (1989).
15. Stanfill, C. and Waltz, D. , Toward Memory-Based Reasoning, CACM, 29-12, (1986), 1213-1228.
16. Sato, S. and Nagao, M. , Toward Memory-Based Translation, Coling '90, (1990).
17. Sadler, V. , Working with Analogical Semantics, Foris Publications, (1989), 256.
18. Nitta, Y. , Idiosyncratic Gap: A Tough Problem to Structure-bound Machine Translation, Proc. of Coling 86, (1986), 107-111.