Information-based Machine Translation

Keiko HORIGUCHI
Spoken Language Technology, Sony US Research Laboratories
3300 Zanker Road
San Jose, CA 95134
keiko@slt.sel.sony.com

Abstract

This paper describes an approach to Machine Translation that places linguistic information at its foundation. The difficulty of translation from English to Japanese is illustrated with data that shows the influence of various linguistic contextual factors. Next, a method for natural language transfer is presented that integrates translation examples (represented as typed feature structures with source-target indices) with linguistic rules and constraints. The method has been implemented, and the results of an evaluation are presented.

Introduction

High-quality automatic translation requires the disambiguation of common, highly ambiguous verbs, such as *to have*, *to take*, or *to get*. It also requires the correct handling of non-compositional, idiomatic expressions with varying degrees of "fixedness". We view Machine Translation in terms of linguistic information represented as typed feature structures. By integrating translation information represented as example pairs with other types of linguistic information represented as rules, our approach extends the capabilities of current machine translation methods, and solves a number of key problems.

1. Linguistic Context for Translation

In translating different words, phrases, and expressions, different types and amounts of information from the context need to be considered. (Only the sentential context is considered here.) So far, a systematic solution to this problem has not been found. This section illustrates the extent of this problem, and the remainder of this paper describes our approach.

1.1. Expressions with *to have*

We examined the problem of translating the English main verb *to have* into Japanese. The verb *to have* was selected because it is quite common in colloquial English, yet forms a large variety of senses, collocations, and idioms. 615 different expressions containing the English verb *to have* were extracted from a 7000-sentence corpus from the "international travel" domain. Each English expression was manually translated into Japanese in the most general way possible.

1.2. Target-language Distinctions

The most general translation for the construction "*X have Y*" in this domain was found to be *Xに Yがある (X-ni Y-ga aru)*:

The copy shop next door has a fax machine.
となりのコピー屋にファックスがあります。
tonari-no kopiiya-ni fakkusu-ga arimasu.
next-ATT copy shop-LOC fax-NOM exist

Other translations are often necessary when the target language imposes finer semantic distinction on the state or on the action that is described. For example, if the object noun phrase refers to one or more human beings, the Japanese verb *aru* is changed into *iru.* Similarly, the word *pet* or a pet animal as the object noun phrase triggers the translation of *to have* as *katteiru,* a Japanese verb for keeping an animal as a pet :

We have two sons.
息子が二人います。
musuko-ga futari imasu.
son-NOM two-CONTR exist

Do you have pets?
あなたはペットを飼っていますか。
anata-wa petto-wo katte-imasu-ka
you-TOP pet-ACC keep-ST-Q

Other examples of finer target-language distinctions include a symptom/disease as the object of *to have*. While many physical symptoms and minor diagnoses (e.g. *pain, cavity, fever, allergy*) use the default translation (*X-ga aru*), a serious illness or diagnosis is translated into the copula construction. Many other *to have* constructions with a symptom/disease object require verbs that are specific to the object noun phrase in Japanese:

I have diabetes.
私は糖尿病です。
watashi-wa toonyobyoo desu.
I-TOP diabetes COP

My wife had a stroke last year.
妻は去年脳卒中で倒れました。
tsuma-wa kyonen noosocchu-de taoremashita
wife-NOM last year stroke-with fall-PST

My husband had a heart attack.
夫が心臓発作を起こしました。
otto-ga shinzoohossa-wo okoshi-mashita
husband-NOM heart attack-ACC cause-PST

1.3. Adjuncts in the Source Language

Some verbal adjuncts can affect the translation of the *to have* construction, not by altering the basic sense of 'existing', but by adding further information to specify the way in which something 'exists'. One example of such an adjunct is a prepositional phrase (PP) whose object noun phrase shares its referent with the SUBJ of *have*. For example, the utterance below expresses that *the map* is held or carried by the speaker, and the Japanese translation uses the verb *motte-iru*, literally meaning *to be carrying/holding*.

I have the map with me.
私はその地図を持っています。
watashi-wa sono chizu-wo motte-imasu.
TOP the map-ACC hold-ST

If the subject noun phrase is inanimate, the Japanese translation uses the verb *tsuite-iru*, which literally means *to be attached.*

The main dish has a salad with it.
メインディッシュにはサラダがついています。
meindisshu-ni-wa sarada-ga tsuite-imasu.
main dish-LOC-TOP salad-NOM attach-ST

Similarly, a construction with an *on*-PP is translated into the Japanese construction *notte-iru*, which literally means *to be written/placed on.* A construction with an *in*-PP is translated into the Japanese construction *haitte-iru*, which literally means *to be placed in*:

Does the map have subway lines on it.
その地図に地下鉄線がのっていますか。
Sono chizu-ni chikatetsusen-ga notte-imasu-ka.
the map-LOC subway line-NOM written-on-Q

The closet has extra hangers in it.
クロゼットに余分のハンガーが入っています。
kurozetto-ni yobun-no hangaa-ga haitte-imasu.
closet-LOC extra-ATT hanger-NOM placed-in-ST

Adjunct adjectival phrases and past participles also specify the way something exists. For example, *available* in the *have* construction generally changes the translation to *aite-iru, to be open or available*:

We have one twin room available.
ツインの部屋が一つ空いています。
tsuin-no heya-ga hitotsu aite-imasu
twin-ATT room-NOM one-CONTR open-ST

1.4. Source Language Ambiguities

In some cases, the *to have* construction in English carries more than one sense, and some linguistic contexts can bring out one of the senses as the preferred meaning. For example, the construction *X has a Y taste* is ambiguous between *to be exercising Y (personal) taste* and *to taste X*. This ambiguity is usually resolved by looking at the semantic properties of the subject noun phrase, as illustrated in the examples below:

He has simple tastes.
彼がシンプルな趣味をしている。
kare-ga shinpuru-na shumi-wo shiteiru
he-NOM simple taste-ACC do-ST

This wine has a very clean taste.
このワインはとても爽やかな味がする。
kono wain-wa totemo sawayaka-na aji-ga suru
this wine-TOP very refreshing taste-NOM do

When the object refers to a specific type of information, such as *number* or *address*, the construction is inherently ambiguous between *to*

know (the number), to be carrying (the number), and *(for the number) to exist.* The construction usually carries the meaning of *to know*, but if the construction is negated, then the sense of *to be carrying* becomes more preferred, since the negative construction is more specific and only negates the proposition that the object is accessible:

I don't have his phone number.
私が彼の電話番号を持っていない。
watashi-ga kare-no denwabangoo-wo motteinai
I-NOM he-GEN phone number-ACC hold-ST-NEG

On the other hand, if the object noun phrase is an indefinite noun phrase, it is more likely to mean *to exist* :

Do you have an extension number?
内線番号がありますか。
naisen bangou-ga arimasu-ka
extension number-NOM exist-Q

Another example of the ambiguities of *to have* concerns the two senses *to have something available* and *to eat,* when the object noun phrase refers to an edible entity. Our corpus analysis shows that some of the linguistic contexts bring out one of the two senses as clearly preferred. For example, the past tense or the perfective aspect brings out the *to eat* sense, whereas the present tense without any aspect markers suppresses this sense:

I had raw fish for dinner.
魚の刺身を夕食に食べました。
sakana-no sashimi-wo yuushoku-ni tabemashita
fish-ATT raw-ACC dinner-GOAL eat-PST

I don't have any American beer on tap.
アメリカの生ビールはありません。
amerika-no namabiiru-wa arimasen.
America-ATT draft beer-TOP exist-NEG

1.5. Support Verb Constructions and Idioms

In some of the constructions, *to have* functions as a support verb. In the support verb construction the object noun phrase constitutes a part of the verbal predicate rather than an argument of the verb. If the target language does not have an equivalent support verb construction, such an expression with a support verb construction has to be translated into the corresponding single verb construction.

Idiomatic expressions in the source and target languages, and their varying degrees of "fixedness", also play a role. For example, the word *見当 (kentoo),* the Japanese translation of *a clue* in *I don't have a clue*, requires the special verb *つく (tsuku),* to constitute an idiomatic expression *見当がつく (kentoo-ga tsuku).* As another example, the English expression *Have a good one* does not allow a compositional translation into a Japanese construction with a main verb plus an object.

1.6. Discussion

From the data described above, it is clear that there are various factors that contribute to the different patterns of translation. In order to handle these different translations correctly, it is necessary to identify the linguistic features of the context that trigger different translations, and to determine how the different features and contexts interact. In the case of the English *to have* construction, the following surface linguistic features are identified that can be interpreted as 'triggers' for translations other than the default translation:

- past tense
- interrogative or imperative constructions
- negative
- modal auxiliaries
- progressive and/or perfective aspect
- adjectival modifiers for the object NP (noun phrase)
- prepositional phrase modifiers for the object NP
- adjectival modifiers for the VP (verb phrase)
- prepositional phrase modifiers for the VP
- adverbial modifiers for the VP
- constructions that carry a pragmatic force (request, suggestion, etc.)

We found that some of the factors have stronger influence on the translation than others. For example, consider the following expression:

Can I have a look at the room?
その部屋を見られますか。
sono heya-wo mi-raremasu-ka.
the room-ACC look-PTN-Q

The source-language expression contains more than one factor that can trigger a different translation. The first factor is the construction that usually carries the pragmatic force of "request", *Can I have X?*, which usually triggers the *X をお願いできますか* (*X-wo o-negai dekimasu-ka)* construction. At the same time, the object noun phrase *a look* means that the verb *to have* is used as a support verb. For this reason, the combination of the verb *have* and the object noun phrase *a look* has to be translated into Japanese as the verbal predicate *見る (miru)*. This shows that the translation preference that is triggered by the root string of the object noun phrase is stronger and should take preference over the translation preference that is triggered by the pragmatic force.

2. Information-based MT

We argue that the sorts of complex translation correspondences that were illustrated in the previous section are best represented as translation examples, but that the transfer procedure must use qualitative linguistic constraints in order to choose the correct examples. Given the types of linguistic features that influence translation, a highly expressive linguistic representation for both input and translation examples is required. We employ typed feature structures throughout all stages of translation.

Since there are complex interactions among different contextual factors, a single quantitative matching function that calculates a distance between the input and the examples is not sufficient. Multiple steps of matching are needed, each considering a small number of linguistic dimensions, with the steps executed in the appropriate order. This is best achieved with a rule-based linguistic transfer procedure that controls the example matching procedure.

2.1. Transfer Component Architecture

The transfer component for information-based MT consists of two main procedures, the linguistic transfer procedure and the example matching procedure. This is illustrated in Figure 1. The input to this component is the source-language typed feature structure; this is created by an analysis component that is not described further here. Similarly, the output of the transfer component is a target-language typed feature structure, from which the target-language expression is generated by the generation component (also not described further).

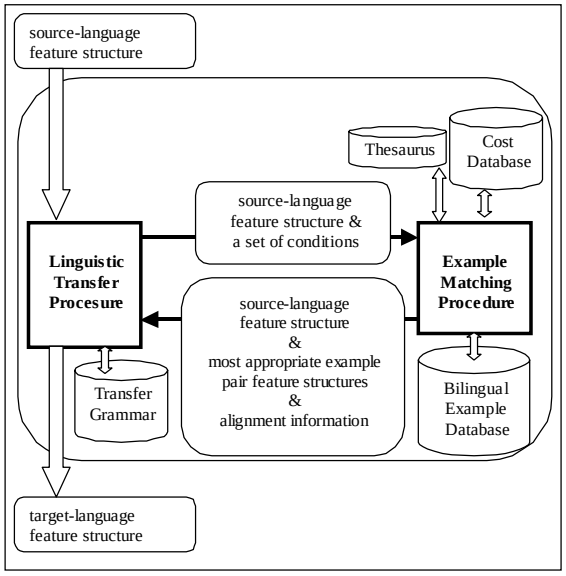

Figure 1: Overview of the Transfer Component

2.2. Linguistic Transfer

The linguistic transfer procedure is implemented as a rewrite-grammar using the special-purpose Grammar Programming Language (GPL) (Duan, et al. 2000, Franz, et al. 2000a). The general role of the transfer grammar is to operate on the input feature structure in a recursive manner, and to perform source-to-target transfer by invoking the example matching procedure, and by using the translation examples to construct a target-language feature structure. The transfer grammar implements the principle of "large to small" in covering the input feature structure.

When the transfer procedure invokes the example matching procedure, it implements the principle of "specific to general". Since the linguistic features interact with each other when they are combined, and since some of the features have more influence on the translation than others, it is necessary to specify a number of separate invocations of the example matching procedure, and to pay particular attention to their order. The invocations of the example matching procedure are arranged so that each call focuses on one or two features, making sure that both the input and the example contain the same feature(s). Different invocations of the matching procedure

are ordered so that the system checks the existence of the most important factors first, gradually progressing to the least important factors.

2.3. Example Matching

The example matching procedure matches the input feature structure against the example feature structures, and it returns the most appropriate example. The architecture of this module is shown in Figure 2.

When the transfer procedure invokes the example matching procedure, it specifies a set of linguistic constraints on which examples may be considered. This is used to narrow down the search space from all the examples to a much smaller set. The examples that satisfy these constraints are matched in detail against the input feature structure. The detailed match is a recursive process operating on the two feature structures that is based on costs for inserting, deleting, or altering features, and on certain constraints for particular features. Lexical similarity is calculated from the thesaurus on the basis of the information content of the thesaurus nodes.

During example matching, the input feature structure is aligned with the example feature structure. The alignment information is used by the transfer procedure to handle differences between the input and the example. For example, if the input contains grammatical features, modifiers, adjuncts, or sub-constituents that are not in the examples, then they are transferred to the target-language representation. Similarly, if the example feature structure contains information that is not present in the input, then the transfer procedure deletes the relevant information.

3. Example Database

The example database contains a large set of translation examples represented as pairs of typed feature structures in the source and target languages. Using a Treebanking tool, the examples are disambiguated, and indices that show corresponding constituents are added. In addition to the type and complexity of the example feature structures, there are three methods for identifying the degree of linguistic

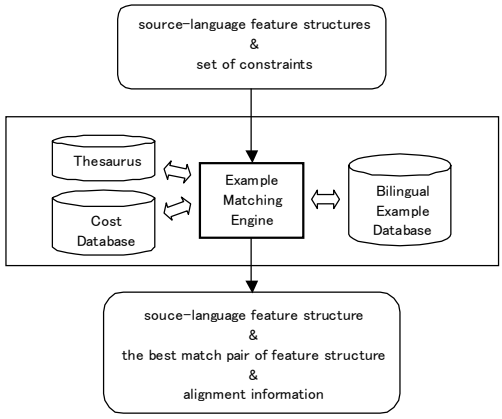

Figure 2: Architecture of the Example Matching Procedure

specificity of an example: marked examples, example indices, and semantic constraints. This information is used by the transfer procedure and the matching procedure to select the best example, using the mechanism of linguistic matching constraints that was described above.

3.1. Marked Examples

Some of the features that were shown in Section 2 to influence the translation have been traditionally described as “marked“. Examples include negation, interrogative, and also the presence of certain adjuncts. The transfer procedure regards these examples as more specific than unmarked examples, and (via the linguistic constraints passed to the matching procedure) only allows such examples when appropriate.

3.2. Example Indices

Examples can contain two types of indices linking a source-language sub-feature-structure with a target-language sub-feature-structure. A CORRESPOND-INDEX signals that the two constituents correspond to each other, while a REPLACE-INDEX signals that two constituents correspond to each other and can be replaced by similar constituents.

The absence of such indices in a major argument phrase (such as the subject or object) indicates that the example is more specific. A CORRESPOND-INDEX is more specific than a REPLACE-INDEX, since a CORRESPOND-INDEX indicates that although the head of the constituent allows modifiers, the constituent can not be substituted. For example, the object *the*

bucket in the example for the idiom *to kick the bucket* does not contain any indices, since the idiom does not allow substitution or modification. On the other hand, *a heart attack* in *to have a heart attack* allows modifiers (e.g. *a severe heart attack*), so the example for the idiomatic translation carries a CORRESPOND-INDEX.

3.3. Semantic Constraints

The example database also contains certain semantic constraints on source-language sub-feature-structures. When an input feature structure is matched with such an example, the matching procedure checks whether the input satisfies the semantic constraint. If it does, then that example is preferred over other examples, since it is more specific than other examples that do not carry a semantic constraint. On the other hand, if the input does not match the constraint, then the match is rejected.

3.4. Sample Entry

Figure 3 shows the example pair for the expressions *Can I have your name?* ⇔お名前をお願いできますか (*o-namae-wo o-negai dekimasu-ka*). This example has a number of marked features. The mood of the sentence is yes-no question, the modal auxiliary *can* is present, and the subject does not contain an index. These features are used by the transfer procedure to ensure that the example is only used to translate appropriate input.

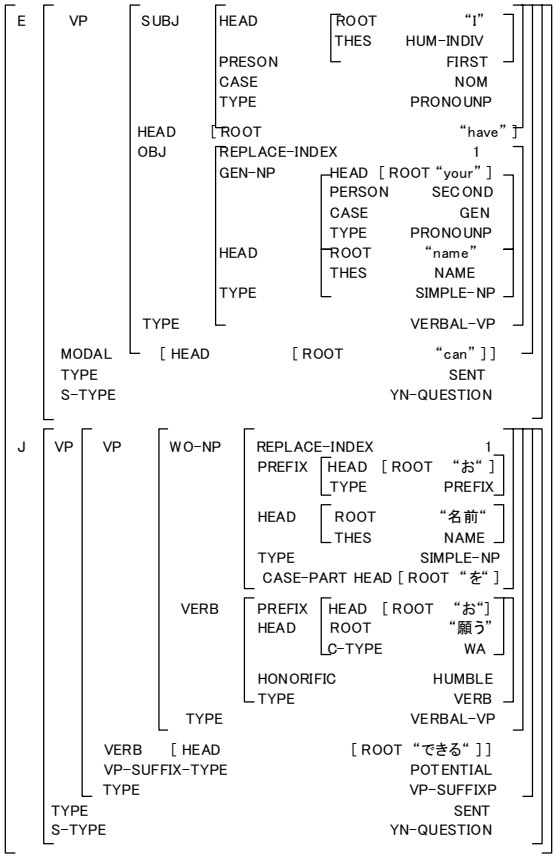

Figure 3: Excerpt from the example database entry for *Can I have your name?*

4. Implementation and Evaluation

A prototype implementation of this translation method has been created by the Sony USRL Speech Translation group (Franz et al. 200b). The prototype was developed for the "overseas travel domain", which includes utterances and expressions useful for travel between e.g. Japan and the USA.

4.1. Lexicon and Example Database

The English-to-Japanese translation system includes an English dictionary with 6483 unique English root forms, and the English-to-Japanese example database contains 14,281 separate example pairs. These entries consist of constructions of various sizes, ranging from conjoined sentences to individual words. For some example pairs, the system automatically extracts corresponding parts from the source and target expressions, and creates a new example pair. As a result, the system has a total of 24,072 example database entries available.

4.2. Development Set

We developed, tested, and refined the system until all of the main predicates of the 615 development set sentences with *to have* were translated correctly. For this, the system used 129 distinct example pairs with the main verb *to have*. Many example pairs encode a specific translation: 68 out of the 129 entries were used to translate only one expression from the development set. On the other hand, some entries are very general, and are used to translate a large number of expressions. The most frequently used entry is *Do you have sushi?* ⇔すしがありますか *(sushi-ga arimasu-ka)*, which is used to translate 113 out of the 615 development set expressions.

4.3. Linguistic Transfer

The transfer grammar contains 153 context-free rules. Each rule includes a rule-body with GPL statements, which can include calls to the example matching procedure, and calls to sub-transfer rules. To translate the 615 expression in the *to have* development set, the system performed an average of 3.4 match-and-transfer steps. (In many cases, more than one transfer path was pursued.) Only 26 out of the 615 expressions were translated with only one match-and-transfer step. Examples of such expressions include *Have a good one!* and *You can have it.* At the other extreme, the maximum number of match-and-transfer steps required to translate a single input expression was 9. One of the expressions that required 9 match-and-transfer steps was *The double on the third floor has a really nice view of the ocean.*

4.4. Evaluation

The system was evaluated using a new corpus of unseen expressions with the verb *to have.* The evaluation data was collected from three different travel phrase books published by Barron, Berlitz, and Lonely Planet. The English expressions containing *to have* as a regular verb (and *have got* as a main predicate) were manually extracted from the phrase books. There were 405 unique expressions with *have* in the resulting evaluation corpus, with an average of 5.5 words. The evaluation corpus was translated by the translation system, and each of the output expressions was examined and manually categorized according to its translation quality. The result is shown in the table below:

.

Flawless Translations	351	86.7%
Incomplete Translations due to OOV	48	11.8%
Wrong Translations	6	1.5%
Total	405	100%

The category "flawless translation" refers to translations without any obvious flaws or problems. "Incomplete translations due to OOV" refers to translations where the main predicate was correctly translated, but due to some out-of-vocabulary (OOV) nouns or modifiers, parts of the source-language input words were carried through to the target language expression. The category "wrong translation" refers to translations where the main predicate is incorrectly translated, with or without out-of-vocabulary words.

4.5. Discussion

Some of the wrong translations are due to ambiguities in the object noun phrase, such as *a fall* in *My child has had a fall*, which the system translated as *watashi-no kodomo-wa aki-ga arimashita* (meaning *My child had an autum).* There were also a number of expressions that should have been translated into different predicates in Japanese, but which were not covered in the example database. Examples of these include the following :

Input: *I've got a nosebleed.*
Output: *鼻血があります*
hanaji-ga ari-masu
nosebleed-NOM exist
Appropriate Translation:
鼻血が出ています
hanaji-ga dete-imasu
nosebleed-NOM come out-ST

The evaluation shows that the information-based translation method works reliably for translating short, single-clause utterances. In support of the generality of this method, we found that translation accuracy could be improved by adding more examples, and that the features that mark specificity of example entries are applicable to expressions with other common verbs besides *have*.

4.6. Future Work

One difficult problem remains in the treatment of support verb constructions. When the object has a modifier, the modifier has to be transferred as a verbal modifier in the target language if the target language requires a single verb construction. For example, *to have a close look* is translated as *to look closely*, and *to have another look* is translated as *to look again.* There are, however, not enough data in the development set to draw any conclusions about how general these modifiers can be treated across different support verb constructions.

One hypothesis is that there are different degrees of proximity between the support verb and the object noun phrase. In some cases, there might be only one fixed phrase to be interpreted as the support verb construction, while other cases may

allow many different modifiers for the object noun phrase. This is suggested by the case of *to have a seat* in the development set. This phrase allows the interpretation of *to sit* only if the object noun phrase is exactly *a seat*. The expression *to have another seat* cannot be translated as *to sit again*, but more like *for another seat to exist*. Further analysis of support verb construction data, including instances with other verbs besides *have*, will be necessary to determine how these constructions can best be handled in the current framework.

Another avenue for future work is the use of Machine Learning techniques to select linguistic features, and statistical methods (such as loglinear models) to model the effect of feature combinations.

Conclusion

The approach described in this paper is based on the conviction that natural language transfer must be driven by qualitative, linguistic information. The analysis of the problem of translating one construction from English to Japanese has shown that a significant amount of linguistic information is necessary for achieving high-quality translation of something as simple as single-clause input. The transfer method that this paper described as one possible solution can integrate translation examples with linguistic rules and constraints in an effective manner.

The linguistic information used in this approach is general and domain-independent; domain-specific translation knowledge is confined to the example database. This modular system architecture presents significant advantages for developing, maintaining, and extending a practical machine translation system.

Acknowledgements

I would like to thank my advisor Prof. Jun'ichi Tsujii, my colleagues at Sony USRL in California, my colleagues at Sony in Tokyo, and the anonymous reviewers of this paper.

References

Duan, L., A. Franz and K. Horiguchi (2000) "Practical Spoken Language Translation Using Compiled Feature Structure Grammars", in *Proceedings of International Conference of Spoken Language Processing (ICSLP-2000)*, Beijing, China.

Franz, A., K. Horiguchi and L. Duan (2000a) "An Imperative Programming Language for Spoken Language Translation", in *Proceedings of International Conference of Spoken Language Processing (ICSLP-2000)*, Beijing, China.

Franz, A., K. Horiguchi, L. Duan, D. Ecker, E. Koontz and K. Uchida (2000b) "An Integrated Architecture for Example-based Translation", in *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, Saabrucken, Germany.

Furuse, O. and H. Iida (1996) "Incremental translation utilizing constituent-boundary patterns", in *Proceedings of COLING-96*, pages 412-417.

Horiguchi, K. (2000) *Integrating Linguistic Information into Example-based Machine Translation*, Ph.D. thesis, University of Manchester Institute of Science and Technology.

Maruyama, H. and H. Watanabe (1992) "Tree cover search algorithm for example-based translation", in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, Montreal, pages 173-185.

Nagao, M. (1984) "A framework of a Machine Translation between Japanese and English by analogy principle", in *Artificial and Human Intelligence*, A. Elithorn and R. Banerji (eds.), North Holland, pages 173—180.

Sato, S. and M. Nagao (1990) "Toward memory-based translation", in *Proceedings of COLING-90*, vol. 3, Helsinki, Finland, pages 247—252.

Sumita, E., O. Furuse, and H. Iida (1993) "An example-based disambiguation of prepositional phrase attachment", in *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-93)*, Kyoto, pages 80-91.

Watanabe, H. (1992) "A similarity-driven transfer system", in *Proceedings of COLING-92*, Nantes, France, pages 770-776.

Watanabe, H. and K. Takeda (1998) "A pattern-based Machine Translation system extended by example-based processing", in *Proceedings of ACL-COLING-98*, pages 1369-1373.