An Idiom-based Approach to Machine Translation

Hagyu Lee and Yung Taek Kim

Department of Computer Engineering
Seoul National University
Seoul 151-742, South Korea

E-mail: kshalt@krsnucc1.bitnet

Abstract

One of the most difficult problems in machine translation is how to deal with those expressions that cannot be translated compositionally from the translations of the parts. From a bilingual point of view, such expressions can be thought of as idioms. And another problem is the ambiguities in analysis. To cope with these problems, we developed an idiom-based approach. In the approach, idiomatic expressions are translated based on idiom recognition and it is performed before parsing to reduce the ambiguities that can be resolved prior to parsing. In the idiom recognition, if an idiom is matched successfully and it is estimated to be safe, the dependencies between the words matched by the idiom are fixed. Therefore, the parser need not attempt to find out alternative dependencies between the words. The safety of an idiom is determined based on the dispersion of its constituents in a sentence. The experimental results show that most of the idioms can be estimated to be safe and that almost all of them, including discontinuous ones, are recognized correctly before parsing by the idiom recognition mechanism.

1 Introduction

In machine translation, it is well known that there are a lot of expressions which cannot be translated by word-to-word translation. The translation of such expressions is one of the focal issues in current machine translation systems. In traditional rule-based transfer approach, the knowledge necessary to the translation of the expressions is encoded in the

complex form of rules. But the development of the rules is a very difficult and time consuming task. It causes a serious problem -- *knowledge acquisition bottleneck* [Santos90, Sato91]. To overcome such problem, we propose an *idiom-based translation* approach where the expressions are translated based on *idiom recognition*.

In this paper, *idioms* are defined as those expressions that cannot be translated compositionally from the translations of the parts [Santos90]. If an expression cannot be translated without structural changes or the changes of dependency relations, it is regarded as an idiomatic expression according to the criterion of the idiom. For example, the expression '감기/에(cold/LOCATION) 걸리/다(be-hung/ENDING)', whose meaning is 'catch cold', is classified into idiom. Because the dependency relation of '감기' is changed from 'LOCATION' to 'OBJECT' in Korean-English translation. In contrast, the expression '약/을(medicine/OBJECT) 먹/다(eat/ENDING)', whose meaning is 'take a medicine', is not regarded as an idiom, though it cannot be translated by the default translations, the translations of the words in general cases, given in the parentheses. Because it can be translated without structural changes or the changes of dependency relations in Korean-English translation. Such expression is called a *collocational expression* [Ok92] and can be translated more easily than idioms. The former example shows the bilingual viewpoint of idioms well, since it is not an idiom in Korean but it belongs to idioms in Korean-English translation.

Through the examination of corpra, it is noticed that the idioms have some good properties. One of the important properties is that idioms show a strong tendency to the localization of their constituents. In other words, when an idiomatic expression appears in a sentence, the possibility that the constituents are dispersed widely in the sentence is very low. This means that most of the idioms including discontinuous ones can be recognized correctly without knowing the global structure of the sentence. For example, long other constituents may intervene between the two words of the idiom '감기/에 걸리/다', but the cases are very rare in real sentences. Another property is that in case of an ambiguity between literal and idiomatic readings, there is clear preference for the idiomatic reading [Linden90]. In consideration of the properties, idioms are recognized prior to parsing in the idiom-based approach, and when an expression is recognized as an idiom, all the other interpretations for the constituents are ignored in parsing and transfer. Thus, not a few ambiguities can be resolved gracefully by the idiom recognition before parsing.

This paper describes an overview of the idiom-based translation approach, especially focusing on the idiom recognition mechanism in Korean-English machine translation system.

2 Idiom-based Machine Translation

Figure 1 shows the basic configuration of the idiom-based translation system. The idiom database, which is a collection of idioms, is the main knowledge source of the system. The idiom recognition component retrieves candidate idioms from the idiom database and identifies those expressions that are matched by the idioms. The parsing component determines the dependencies not identified by the idioms and produces a source language dependency tree after ambiguity resolution. In tranasfer, the target expressions of the idiomatic constituents and nonidiomatic constituents are composed to fit for the target language. The target expressions of the idiomatic constituents are given by the idiom recognition component and the target expressions of the nonidiomatic constituents are chosen in this component from the bilingual dictionary. During the composition, the matching relations of the idioms, which are produced in the idiom recognition, are helpfully used. The target language dependency tree, the result of the transfer component, is then passed to generation component, not shown here.

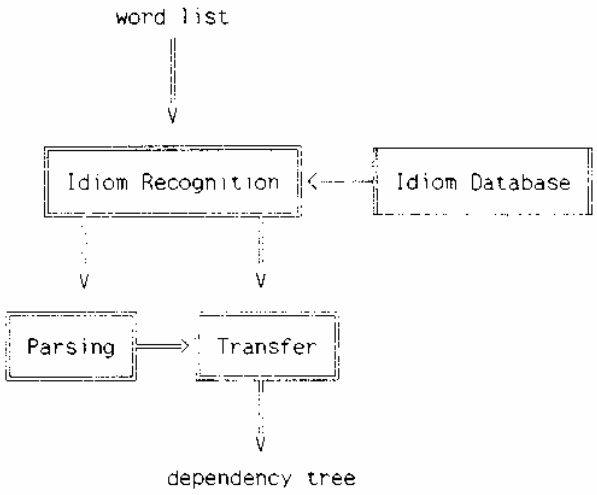

Figure 1. Basic configuration of idiom-based translation

3 Representation of Idiom

The idiom database contains Korean-English bilingual idioms. Every idiom is stored in the idiom database so that it can be retrieved by its main word and part of speech. To recognize idioms correctly, especially before parsing, and to give precise translations for the idiomatic expressions, the representation of idioms is important.

An idiom (ID) consists of a source expression (SE), a target expression (TE), a set of special constraints (CS) and a set of correspondence links (CL).

(1) ID = <SE, TE, CS, CL>

Two examples are shown below. The SE of (2) is not an idiomatic expression from a monolingual point of view in Korean but the SE of (3) is an idiomatic expression in Korean, too.

```
                  1       2                ; position number
(2) <[[감기/에] 걸리/-],                   ; [[cold/LOCATION]  be-hung]
                  1       2                ; => catch cold
     [catch/v [cold/OBJ,n]],
     {},
     {(1,2), (2,1)}>

(3) <[[[A/ㄹ ] 지/를] 모르/-],             : [[[A/RELATIVE] bound-noun/OBJECT] don't-know]
    [A/MAY],                               : => may A
    {AJC(1,2), AJC(2,3)},
    {(1,1)}>
```

As shown above, the SE is a Korean dependency tree of idiom units (IUs) for an idiomatic expression. An IU consists of a base part and a suffix part. The base part describes the base form of an IU. It may be either a lexically fixed form or a variable. The former is called a "constant base", the latter a "variable base". Every ID has at least one constant base IU. The variable base can match any words if no special constraint is given for it. The suffix part represents the dependency relation between an IU and its governor IU by the representative suffix form. Hereafter, the term "suffix" will be used to indicate the inflectional suffix, such as Korean postposition or ending, which represents grammatical relation. Korean is a head-final language. Therefore, all the dependents come before their governor in a sentence. It is the same in SE, so the governor IU comes last

in a local tree of a SE. Also, it is common that there is no ordering restriction between the dependents even though they belong to idioms, since Korean is a partially-free word order language. Therefore, no linear precedence is assumed between the dependents if no ordering restriction is given in CS.

The TE is an English dependency tree of target units (TUs) corresponding to SE. A TU consists of a base part and an information part. The form of the base part is the same as that of the SE. The information part can contain the information necessary to generation such as dependency relations, parts of speech, etc. In contrast to SE, TE is ordered between the dependents so as to make the word ordering easy in generation. From now on, the notations IU(i) and TU(j) will be used to indicate the *i*th IU and the *j*th TU, respectively.

Special constraints that the SE must satisfy in idiom matching are specified in CS. Linear precedence, adjacency condition and constraints on variable bases can be included in the CS. Some idioms require the word ordering between the dependents. A linear precedence is specified in the form of LP(i,j) which means that IU(i) precedes IU(j). There are a lot of idioms that allow other constituents to intervene between the idiomatic constituents, since the definition of idioms is extended to the bilingual one in the idiom-based approach. So, the intervention is assumed to be allowed if no adjacency condition is given. An adjacency condition is specified in the form of AJC(i,j) which means that IU(i) and IU(j) are adjacent. The constraint on the variable base takes the form of *predicate*(i). When it is specified, the predicate must be true for the word matched by the variable base of IU(i) if the idiom is to succeed in matching.

CL contains correspondence links. A correspondence link is a pair of positions (i,j) which represents that there is a correspondence from IU(i) to TU(j) in transfer.

4 Idiom Recognition

Figure 2 shows the flow of idiom recognition. Idioms are recognized passing through three steps. First, candidate idioms are retrieved from the idiom database. Then, it is determined whether they match the word list or not. And finally, idioms that are consistent with each other are selected from the matched idioms. After the selection, four results are produced.

Of the results, the selected word list and a set of idiomatic relations of the selected idioms (IR) are passed to the parsing component. The selected idioms and nonlocal idioms are passed to the transfer component. They will be explained later in detail.

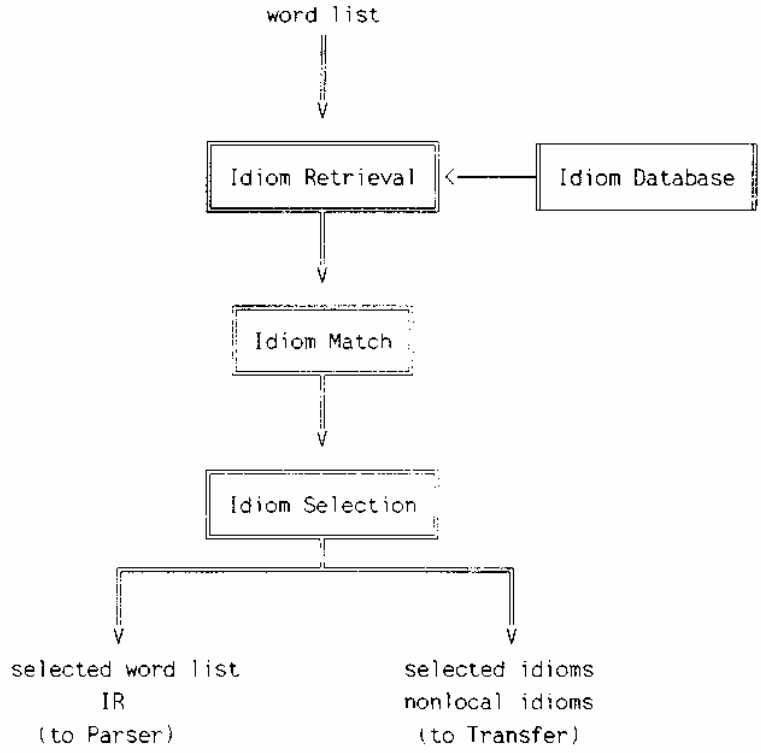

Figure 2. Flow of idiom recognition

4.1 Idiom Retrieval

The input to the idiom retrieval step is a word list which is the result of the morphological analysis of a sentence. If a word has lexical ambiguities, there are more than one lexical tokens in it. From now on, the notations WORD(i) and TOKEN(i,j) will be used to indicate the *i*th word and the *j*th lexical token of the *i*th word, respectively. In this step, all the candidate idioms are retrieved from the idiom database with the keys composed of the base forms and their parts of speech of the lexical tokens. Here is an example of the word list.

(4) input sentence

그녀는 심한 감기에 자주 걸릴 지도 모른다.

: She may catch bad cold frequently.

word list

words		lexical tokens	
1: 그녀는	=>	1: 그녀(n)/는(ap)	; she
2: 심한	=>	1: 심하(v)/ㄴ (e)	; severe, bad(cath a bad cold)
3: 감기에	=>	1: 감(v)/기(nmz)/에(e)	; wind
		2: 감(v)/가에(e)	; wind
		3: 감기(n)/에(cp)	; cold
4: 자주	=>	1: 자주(n)	; purple color
		2: 자주(ad)	; frequently
5: 걸릴	=>	1: 걸리(v)/ㄹ (e)	; be hung
6: 지도	=>	1: 지도(n)	; map
		2: 지(n)/도(ap)	; bound noun
7: 모른다	=>	1: 모르(v)/ㄴ (pf)/다(e)	; don't know

where, n: noun, v: verb, ad: adverb,
cp: case postposition, ap: auxliliary postposition,
e: ending, pf:pre-final ending, nmz: nominalizer

4.2 Idiom Match

In the idiom match step, the idioms that match the word list are chosen from the candidate idioms. The match between an idiom and the word list is considered to be successful if every IU of the idiom matches a lexical token for the base part and suffix part, there is no structural contradiction between the IUs and corresponding lexical tokens and all the special constraints in CS are satisfied. A candidate idiom may match successfully to be related with several possible instances in a sentence. A matched idiom (MID), or an instance of an idiom, consists of the original idiom ID, a set of idiom matching relations (MR) and a set of dependency relations (DR).

(5) MID = <ID, MR, DR>

MR contains matching relations between idiom units and lexical tokens. The matching relation (i,j,k) represents that UI(i) in ID has matched TOKEN(j,k). DR contains the dependency relations between the words matched by the idiom. The dependency relation (i,j,r) represents that there is a dependency relation r between WORD(i) and WORD(j) which is identified by the idioms. For example, (6) and (7) are the MIDs of (2) and (3), respectively, for the word list shown in (4).

(6) <ID, {(1,3,3), (2,5,1)}, {(3,5,에)}>
where, ID = (2)

(7) <ID, {(1,5,1), (2,6,2), (3,7,1)}, {(5,6,ㄹ), (6,7,를)}>
where, ID = (3)

4.3 Idiom Selection

In selection of idioms, we must consider two points:

1. Recognition of idioms prior to parsing.
2. Selection of the best idioms.

If the words matched by an idiom is dispersed too wide in a sentence, it is dangerous to select the idiom, since the possibility that the idiom is recognized incorrectly is comparatively high. Such idioms should not be selected in the idiom selection step. Because the idiom recognition is performed prior to parsing and the global structure of the sentence is not known yet. The safety of an idiom is determined based on the dispersion of the constituents of the idiom. Also, the score of an idiom is introduced to select the best one among candidates. They are explained first, then the selection process will be described.

4.3.1 Dispersion of Idiom

Before defining the dispersion of an idiom, we define the size and range of idiom first. The size of a MID is defined as the number of idiom units in the ID of the MID.

(8) size(MID) = the number of IUs in ID of MID

The range of a MID is defined to be:

(9) range(MID) = j - i + 1
where, MID=<ID, MR, DR>,
i is the minimum value such that (i,k,r)∈DR and
j is the maximum value such that (l,j,s)∈DR.

Then the dispersion of a MID is defined as follows.

$$(10)\quad \text{dispersion(MID)} = \frac{\text{range(MID)}}{\text{size(MID)}}$$

The safety of an idiom is determined by the "dispersion limit value". If the dispersion of a MID is not greater than the dispersion limit value, the MID is said to be "locally matched". Locally matched idioms are selected in the idiom selection step, since the possibility that they are recognized incorrectly is very low. But the selection of nonlocal idioms is postponed till transfer.

4.3.2 Score of Idiom

The score of an idiom should reflect the possibility of correct translation. We calculate the score in consideration of the three main factors:

1. Size of the idioms.
2. Size of the constant base IUs.
3. Dispersion of the idioms.

In general, the larger the translation unit, the more accurate the translation. If there are more constant base IUs, the translation can be more specific and precise. Therefore, the idiom with more IUs, with more constant base IUs and with less dispersion is preferred in the idiom selection.

The constant base size of a MID is defined as the number of constant base idiom units in the ID of the MID.

$$(11)\quad \text{csize(MID)} = \text{the number of constant base IUs in ID of MID}$$

Then the score of a MID is defined as follows.

$$(12)\quad \text{score(MID)} = \frac{\text{size(MID)} \times \text{csize(MID)}}{\text{dispersion(MID)}}$$

The following shows the scores of (6) and (7).

(13)

	dispersion	score
(6)	1.500	2.667
(7)	1.000	6.000

4.3.3 Selection Process

In the idiom selection step, a score is assigned to every MID first. Then they are divided into local idioms and nonlocal idioms according to the dispersion value. The selection of nonlocal idioms is postponed till transfer as mentioned above. And then the selection of local idioms is performed. First, all the consistent subsets of MIDs are generated, then the subset whose sum of scores is maximal is selected. If a set of MIDs satisfies the dependency constraint and the lexical token selection constraint, it is said to be "consistent". The constraints are described as follows.

(14) Dependency Constraint
For a set of MIDs,
if d1 and d2 are dependency relations in different MIDs,
then they must satisfy the following two conditions.
(a) Uniqueness of Governor
If d1=(i,k,r), d2=(j,l,s), and i=j,
then k=l and r=s.
(b) No-crossed Dependency
If d1=(i,k,r) and d2=(j,l,s),
then they are one of the followings.
① $k \leq j$, or
② $l \leq i$, or
③ $i < j$ and $l \leq k$, or
④ $j < i$ and $k \leq l$.

(15) Lexical Token Selection Constraint
For a set of MIDs,
when m1 and m2 are idiom matching relations in different MIDs,
if m1=(i,k,m) and m2=(j,K,n), then m must be equal to n.

The dependency constraint comes from the general characteristics of dependency relations in Korean. The lexical token selection constraint specifies that, in a word, only one lexical token can be matched by the selected idioms. After the selection of local idioms, token selection is performed. If a lexical token in a word is matched by the selected idioms, all the other tokens in the word are pruned. The result is a "selected word list". For example, (16) is the selected word list for the word list (4).

```
(16)  selected word list
      1: 그녀는   =>  1: 그녀(n)/는(ap)          ; she
      2: 심한     =>  1: 심하(v)/ㄴ(e)           ; severe, bad(cath a bad cold)
      3: 감기에   =>  1: 감기(n)/에(cp)          ; cold
      4: 자주     =>  1: 자주(n)                 ; purple color
                      2: 자주(ad)                ; frequently
      5: 걸릴     =>  1: 걸리(v)/ㄹ(e)           ; be hung
      6: 지도     =>  1: 지(n)/도(ap)            ; bound noun
      7: 모른다   =>  1: 모르(v)/ㄴ(pf)/다(e)    ; don't know
```

Finally, a set of idiomatic relations of the selected idioms (IR) is built. An IR consists of the dependency relations that are fixed by the selected idioms and the relations that are proved to be impossible due to the fixed dependency relations. Each idiomatic relation is either of the form (i,j,r) or (i,j,F). (i,j,r) means that r has been fixed upon as the dependency relation between the WORD(i) and WORD(j). (i,j,F) means that the dependency between WORD(i) and WORD(j) is not allowed or a failure. For example, (17) is the IR of the set of MIDs (6) and (7) for the word list (4).

```
(17)  IR = {(3,5,에), (5,6,ㄹ), (6,7,를),
            (3,4,F), (3,6,F), (3,7,F), (5,7,F),
            (4,6,F), (4,7,F), (1,4,F), (2,4,F)}
```

The selected word list and IR are passed to the parsing component. They are used in determining the dependencies that are not identified in this step. By using the selected word list instead of the original word list, the parser are relieved of lexical ambiguity resolution to some extent. And the IR plays an import role in reducing the ambiguities of the dependencies. The selected MIDs and nonlocal idioms are passed to the transfer component. They are used in the composition of a proper target language dependency structure.

5 Parsing

To construct source language dependency trees, the parser applies dependency relation rules developed according to the Korean dependency grammar [Yoon90]. The parsing algorithm is basically the same as the tabular chart version of the Cocke-Younger-Kasami (CYK) algorithm [Nijholt90] except that the entries of the parsing table are filled with not phrase structure trees but dependency structure trees, since the parsing rules are not phrase structure rules but dependency relation rules.

If simplified, the form of a parsing rule is as follows. (Note: The unary rules, used only in initialization of the chart table, will not be explained.)

(18) $\gamma(\alpha,\beta)$

What the rule means is that when two wordforms α and β are given, β can govern α with dependency relation γ. When the length of the word list is n, the parsing table is an upper-triangular (n+1)x(n+1) matrix. Each table entry $t_{ij}(i<j)$ contains the partial trees from WORD(i+1) to WORD(j). The final trees belong to $t_{0,n}$. First, every $t_{i,i+1}$ is initialized as the single node trees constructed from the tokens of WORD(i+1) according to the unary rules. Then, $t_{ij}(i+1<j)$ is filled with the trees that are constructed from the trees in $t_{i,k}(i<k)$ and $t_{kj}(k<j)$. In this step, the idiomatic relations in the IR are referenced. It is performed as follows.

(19) For every d1 and d2 such that d1$\in t_{i,k}$, d2$\in t_{k,j}$ $(i<k<j)$,
(a) if $(k,j,r)\in$IR and $r\neq$F
then construct a tree where β governs α with dependency relation r
and add it to $t_{i,j}$
(b) if $(k,j,r)\notin$IR and there is a rule $\gamma(\alpha,\beta)$
then construct a tree where β governs α with dependency relation γ
and add it to $t_{i,j}$
where, α and β are the top-most nodes of d1 and d2, respectively.

As described above, if idiomatic relations are given in the IR, no other interpretation is made by the parser. If the idiomatic relation is a failure F, no new tree is constructed. Otherwise, a new tree is constructed only for the dependency relation. Thus, the number of ambiguous trees can be reduced greately due to the idiom recognition. As a result, the parsing performance is greately enhanced in its speed and accuracy. If there are more than one trees, disambiguation is performed and only one dependency structure tree is passed to the transfer component.

6 Transfer and Generation

In transfer, the selection of the nonlocal idioms are performed first. Then, the target expressions of the nonidiomatic constituents are chosen from the bilingual dictionary by means of the collocational translation and default translation. And finally, a proper target language dependency tree is composed. The selection of nonlocal idioms is very easy

compared with that of local idioms, since the global structure of the sentence has already been determined. In the composition, the idom matching relations are helpfully used. In generation, the word ordering is facilitated by the target expressions of the idiomatic constituents, since they are ordered as explained in Section 3.

7 Experiment

This system has been developed for the purpose of providing commercial usage. Currently, there are about 80,000 idioms in the idiom database. Table 1 shows the result of comparison test. The test was performed to measure the effect of the idiom recognition. 1,000 test sentences were chosen randomly from technical manuals. According to the criterion of the idiom, 2,051 idiomatic expressions are found in the sentences. The dispersion limit value has been set as 2.5 empirically in the test. Each table entry of row 2 and 3 represents the ratio of the result of the test with the idiom recognition to that of the test without it. The test has shown some valuable results. The number of partial trees are greately reduced due to the idiom recognition. The effect of token pruning is not so great but fixing the dependencies before parsing is very effective. Moreover, the longer the sentences, the more effective the idiom recognition. Considering that the long sentences are hard to deal with in parsing, the result is very impressive.

Sentence Length (# of sentences)	1-5 (232)	6-10 (459)	11-15 (235)	16- (74)	Total (1000)
Lexical Tokens	0.964	0.926	0.924	0.916	0.934
Partial Trees	0.912	0.738	0.594	0.461	0.724

Table 1. The result of comparison test

Another result shows that more than 95% of the idioms are estimated to be local when the dispersion limit value is 2.5, and that about 2% of them are recognized incorrectly. The errors in the idiom recognition caused the parser to fail in parsing 17 sentences, mainly due to the token pruning. When the dispersion limit value is set as 3.0, more than 97% of the idioms are estimated to be local but about 4% of them are not

recognized correctly.

8 Conclusion

In this paper, idioms are defined in a broad sense from a bilingual point of view. This paper presents the idea of idiom-based translation, that is, translation of idiomatic expressions based on idiom recognition. The representation of the idioms and the mechanism that recognizes the idioms before parsing are explained in detail. The brief explanations of the parsing and transfer mechanisms in an idiom-based framework are also given.

The experiment shows very promising results. Most of the idioms are local, so can be selected before parsing. And almost all of the local idioms are selected correctly by means of the idiom recognition mechanism. Therefore, in most cases, not only the best translation can be given by this approach, but also the parsing performance is considerably enhanced due to the disambiguation by the idioms selected before the parsing. Moreover, the longer a sentence, the more effective the idiom recognition.

If we increase the dispersion limit value to enlarge the coverage of local idioms, naturally the accuracy of the idiom recognition decreases. To overcome this problem, we are currently developing a new idiom selection method based on statistical approach, where the frequencies of idioms and part of speech tagging method are used. The preliminary test of the new method shows that the accuracy of idiom recognition can increase 2-3% when the dispersion limit value is 3.

References

[Abeille'89] A. Abeille' and Y. Schabes, "Parsing Idioms in Lexicalized TAGs," *Proc. of 4th Conference of the European Chapter of the ACL*, pp. 1-9, 1989.

[Fukumoto90] F. Fukumoto and H. Sano, "A Framework for Restricted Dependency Grammar," *Proc. of the Seoul International Conference on Natural Language Processing(SICONLP-90)*, pp. 11-16, 1990.

[Linden90] E. van der Linden and W. Kraaij, "Ambiguity resolution and the retrieval of idioms: two approaches," *Proc. of COLING-90*, Vol. 2, pp. 245-249, 1990.

[Nijholt90] A. Nijholt, "The CYK-Approach to Serial and Parallel Parsing," *Proc. of SICONLP-90*, pp. 144-155, 1990.

[Ok92] C. Ok, "A Selection of Best Translation Using Collocation," *Proc. of the 2nd Pacific Rim International Conference on Artificial Intelligence(PRICAI-92)*, Vol. 1, pp. 226-232, 1992.

[Santos90] D. Santos, "Lexical gaps and idioms in machine translation," *Proc. of COLING-90*, Vol. 2, pp. 330-335, 1990.

[Sato91] S. Sato, "Example-Based Translation Approach," *Proc. of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing(FGNLP)*, pp. 1-16, 1991.

[Watanabe92] H. Watanabe, "A Similarty-Driven Transfer System," *Proc. of COLING-92*, Vol. 2, pp. 770-776, 1992.

[Yoon90] D. H. Yoon, "A Parsing Mechanism using Cell Elimination in Dependency Structure Grammar," *Proc. of SICONLP-90*, pp. 156-161, 1990.

[Yoon92] S. H. Yoon, "Idiomatical and Collocational Approach to Machine Translation," *Proc. of PRICAI-92*, Vol. 1, pp. 49-53, 1992.

[Zuijlen89] J. M. van Zuijlen, "Probabilistic Method in Dependency Grammar Parsing," *Proc. of International Workshop on Parsing Technologies*, pp 142-151, 1989.