TRANSLATION AMBIGUITY RESOLUTION BASED ON TEXT CORPORA OF SOURCE AND TARGET LANGUAGES

Shinichi DOI and Kazunori MURAKI
NEC Corp. C&C Information Technology Research Laboratories
4-1-1, Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN
e-mail: doi%mtl.cl.nec.co.jp@sj.nec.com

ABSTRACT

We propose a new method to resolve ambiguity in translation and meaning interpretation using linguistic statistics extracted from dual corpora of source and target languages in addition to the logical restrictions described on dictionary and grammar rules for ambiguity resolution. It provides reasonable criteria for determining a suitable equivalent translation or meaning by making the dependency relation in the source language be reflected in the translated text. The method can be tractable because the required statistics can be computed semi-automatically in advance from a source language corpus and a target language corpus, while an ordinal corpus-based translation method needs a large volume of bilingual corpus of strict pairs of a sentence and its translation. Moreover, it also provides the means to compute the linguistic statistics on the pairs of meaning expressions.

1 Introduction

Recently many kinds of natural language processing systems like machine translation systems have been developed and put into practical use, but ambiguity resolution in translation and meaning interpretation is still the primary issue in such systems. These systems have conventionally adopted a rule-based disambiguation method, using linguistic restrictions described logically in dictionary and grammar to select the suitable equivalent translation and meaning. Generally speaking, it is impossible to provide all the restrictions systematically in advance. Furthermore, such machine translation systems have suffered from inability to select the most suitable equivalent translation if the input expression meets two or more restrictions, and have difficulty in accepting any input expression that meets no restrictions.

In order to overcome these difficulties, following methods are proposed these years:

1. Example-Based Translation : the method based on translation examples (pairs of source text and its translation) [Nagao 84, Sato 90, Sumita 90]
2. Statistics-Based Translation : the method using statistical or probabilistic information extracted from a bilingual corpus [Brown 90, Nomiyama 91]

Still, each of them has inherent problems and is insufficient for ambiguity resolution. For example, either an example-based translation method or a statistics-based translation method needs a large-scale database of translation examples, and it is difficult to collect an adequate amount of a bilingual corpus.

In this paper, we propose a new method to select the suitable equivalent translation using the statistical data extracted independently from source and target language texts [Muraki 91]. The statistical data used here is linguistic statistics representing the dependency degree on the pairs of expressions in each text, especially statistics for co-occurrence, i.e., how frequently the expressions co-occur in the same sentence, the same paragraph or the same chapter of each text. The dependency relation in the source language is reflected in the translated text through bilingual dictionary by selecting the equivalent translation which maximizes both statistics for co-occurrence in the source and target language text. Moreover, the method also provides the means to compute the linguistic statistics on the pairs of meaning expressions. We call this method for equivalent translation and meaning selection DMAX Criteria (Double Maximize Criteria based on Dual Corpora).

First, we make comments on the characteristics and the limits of the conventional methods of ambiguity resolution in translation and meaning interpretation in the second section. Next, we describe the details of DMAX Criteria for equivalent translation selection in the third section. And last, we explain the means to compute the linguistic statistics on the pairs of meaning expressions.

2 Conventional Methods of Ambiguity Resolution

2.1 Rule-Based Translation

In conventional methods, linguistic restrictions described in the dictionary and grammar are used to select the suitable equivalent translation or meaning. In general, these restrictions are described logically on characteristics of another expression which modifies or is modified by the expression to be processed. For example, to translate predicates (verbs and predicative adjectives), semantic restrictions are described on essential case arguments in forms of semantic markers to indicate features of words or terms in the thesaurus to show a hierarchy composed of word concepts.

Though these conventional methods have been very useful to realize natural language processing systems, they have the following problems:

1. It is impossible to decide the most suitable equivalent translation if the input expression meets two or more restrictions.
2. Analysis fails when the input expression can meet no restrictions.
3. Actually the practical systems depends on such heuristics as pre-decided application order of restrictions or some default equivalent translations or meanings.
4. The description of the restrictions is based on direct structural dependencies, therefore it is quite difficult to describe the restrictions based on sister-dependency or between expressions belong to different sentences or paragraphs.
5. Restrictions on any dependencies cannot be thoroughly described in advance.

For example, a Japanese word "*booru*" has two meanings, one is 'a ball(*a round object used in a game or sport*)' and the other is 'a bowl(*a deep round container open at the top especially used in cooking*)'. When this word occurs in the following sentence, it must mean 'a bowl'.

JAP:	*Booru–ni*		*mizu–o*		*ireru*
	bowl or ball	dative marker	water	obj. marker	pour, put in or fill
			⇓		
ENG:	`To pour water into a bowl`				

In this case, to select the meaning by the logical restrictions on dependencies, it is necessary to have described even the appearance or usage of the indirect object of the verb "*ireru*". To describe such detail restrictions on all expressions may be possible, but it is quite difficult because the trouble of description and the cost of calculation.

2.2 Example-Based Translation

Besides the conventional translation method above, a machine translation system based on translation examples (pairs of source texts and their translations) is also proposed [Nagao 84, Sato 90, Sumita 90]. This type of system, called Example-Based Machine Translation, has stored a large amount of bilingual translation examples as a database, and translates input expressions by retrieving an example most similar to the input from the database. There is no failure of output in this method because it selects the most similar example not the identical one.

However this example-based translation system needs a large-scale database of translation examples, and it is difficult to collect an adequate amount of bilingual corpora. Even if it is possible, there is no means to divide the sentences of such corpora into fragments and link them automatically, and it costs us too much time and money to divide and link manually. Besides, this method can neither achieve precise meaning interpretation because it selects equivalent translation directly from the input expression and leaves meaning interpretation out of consideration.

To overcome this problem, we have also proposed a new mechanism based on sentential examples in dictionary, which utilize the merits of both the translation by logical restrictions and the example-based method, by selecting the equivalent translation which has the most similar example to the input expression [Doi 92]. This mechanism can guarantee no failure in selecting an equivalent translation, but the description of relations are still based only on direct structural dependencies.

2.3 Statistics-Based Translation

Several new methods especially of machine translation have been proposed lately, which select a suitable equivalent translation using statistical or probabilistic information extracted from language text [Brown 90, Nomiyama 91]. Because many machine readable texts have been already collected nowadays, it is not difficult to extract statistical information of each expression in the texts semi-automatically. Moreover, the statistical information reflects the context in which each word occurs and implies the logical restrictions based on indirect structural dependencies.

Although we call the systems in a same word "statistics-based translation", statistical information used in the methods is diverse, such as translation probability, connectivity of words, statistics for (co-)occurrence, etc. We make comments on the characteristics and the limits of these systems.

The first method uses fertility probabilities, translation probabilities and distortion probabilities [Brown 90]. Fertility means the number of the words in target language that the word of the

source language produces, and distortion means the distance between the position of the word of the source language and the one of the target language. The method has been applied to an experimental translation system from French to English. However, since these probabilities are extracted from a large amount of text pairs that are translations of each other, this method must be suffered from the same difficulties as example-based translation in collecting and analyzing an adequate amount of bilingual corpora, and it's very difficult to apply this method to the languages whose linguistic structures aren't similar each other, such as English and Japanese.

The second method uses the statistics for occurrence in target language text [Nomiyama 91]. It is calculated in advance how frequently the each expression occurs in the target language text, which needs only to belong the same field as the source language text belongs, but not to be a translated text of the source language text. If there are more than one possible equivalent translations, the most frequent translation is selected through this calculated data. Moreover, this method can be applied to make good use of the conventional methods of selecting equivalent translations, for it employs the frequency data exclusively when logical restrictions cannot select one out of candidates.

However this method has one big problem. The high frequency of the expression in the target language text may not originate from the frequency of the expression in the source language text to be translated, because one target language expression does not correspond to only one source language expression in general.

Suppose the following sentence is a first example:

JAP:	*Sono*	*saibankan–wa*		*kooto–to*
	that	judge	subj. marker	coat and or court
		nekutai–o		*katta.*
		tie	obj. marker	bought
		⇓		
ENG:	`The judge bought a coat and a tie.`			

Figure 1 indicates translation process through bilingual dictionary and the statistics for co-occurrence of each pair of expressions in both Japanese and English necessary to translate the sentence[1]. The Japanese word "*kooto*" has two equivalent English translations: '(over)coat' and '(tennis) court'. We cannot decide which is eligible with only logical restrictions on the direct object of the Japanese verb "*kau*", because we can buy both 'coat' and 'court'—the sentence "*Tenisu–kooto o kau*" = 'To buy a tennis court' is also quite acceptable. In this case, the statistics for co-occurrence in the target language English text denotes that the most frequent pair is 'court–judge', because the word 'court' also means a 'law court'. Then using only statistical data on the target language text misleads a wrong expression 'court' as the equivalent translation of "*kooto*", and the example sentence may be translated into 'The judge bought a court and a tie.'.

The second example is this sentence[2]:

JAP:	*Kotori–no*	*kago–ni*		*mizu–o*
	bird of	cage or basket	dative marker	water obj. marker
	ireta	*booru–o*		*oita.*
	filled	bowl or ball	obj. marker	put
		⇓		
ENG:	`I put a bowl filled with water in the bird cage.`			

Translation process of this sentence and the statistics for co-occurrence are shown in Figure 2. Because the pair of 'basket' and 'ball' co-occurs most frequently in the target language, the sentence may be translated into 'I put a ball filled with water in the bird basket.'.

3 Equivalent Translation Selection by Statistical Data on Dual Corpora of Source and Target Languages

Now we propose a new method to provide reasonable criteria for selecting a suitable equivalent translation or meaning using the simple statistical data extracted from source language text in addition to the one from target language text. These source and target language texts don't have to be translations of each other. The proposed method gives us a way to select the expression with the highest frequency of the target language text that keeps high frequency of the source language text at the same time, so it overcomes the difficulty of the method using the frequency data on the target language text only, because it does not select the expression with the highest frequency of only the target language text.

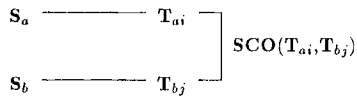
[1]The statistics for co-occurrence of expressions shown in the figures are given provisionally for understanding.

[2]The subject phrase "*watashi–wa*" = 'I' is omitted in this sentence.

3.1 Using statistical data on source language text

The method using only statistical data on the target language text may mislead a wrong equivalent translation, because in general each target language expression corresponds to more than one source language expression.

The equivalent translation selection with statistics for co-occurrence in the target language text when a source language expression $\mathbf{S}_a$ has n equivalent translations in target language $\mathbf{T}_{ai}(i = 1 \cdots n)$ is shown as this:



where

$\mathbf{S}_k$	: source language expression
$\mathbf{T}_{ki}$ $(i = 1 \cdots n)$	: n target language equivalent translations of $\mathbf{S}_k$
$\mathbf{SCO}(\mathbf{E}_i,\mathbf{E}_j)$	: statistics for co-occurrence of two expressions $\mathbf{E}_i,\mathbf{E}_j$

The method using only statistical data on the target language text selects $\mathbf{T}_{ai}$ which maximizes the statistics for co-occurrence in the target language text [3] as the equivalent translation of $\mathbf{S}_a$, where the partner of the co-occurrence $\mathbf{T}_{bj}$ plays the part of the basis for the equivalent translation selection. The biggest problem of this method is that $\mathbf{T}_{bj}$ which depends both b and j is selected by only statistical data on the target language text.

Our new method provides reasonable criteria for selecting the basis for the equivalent translation selection using the statistical data on the source language text. First the source language expression $\mathbf{S}_b$ which maximizes the statistics for co-occurrence in the source language text [4] is selected, then the equivalent translation $\mathbf{T}_{ai}$ which maximizes the statistics for co-occurrence in the target language text [5] is selected. The dependency relation in the source language is reflected in the translated text through this method. We call this method for equivalent translation and meaning selection DMAX Criteria (Double Maximize Criteria based on Dual Corpora).

3.2 Double Maximum Criteria based on Dual Corpora

The algorithm of this method is summarized as follows:

1. Prepare the source and target language texts (the target language text needs not to be a translated text of the source language text).
2. Accumulate the statistics for co-occurrence of every expression in both texts.
3. When a source language expression $\mathbf{S}_a$ has n equivalent translations in target language $\mathbf{T}_{ai}(i = 1 \cdots n)$
   (a) Select $\mathbf{S}_b | \max_b \mathbf{SCO}(\mathbf{S}_a, \mathbf{S}_b)$
   (b) Select $\mathbf{T}_{ai} | \max_{i,j} \mathbf{SCO}(\mathbf{T}_{ai}, \mathbf{T}_{bj})$

[3] $\mathbf{T}_{ai} | \max_{b,i,j} \mathbf{SCO}(\mathbf{T}_{ai}, \mathbf{T}_{bj})$
[4] $\mathbf{S}_b | \max_b \mathbf{SCO}(\mathbf{S}_a, \mathbf{S}_b)$
[5] $\mathbf{T}_{ai} | \max_{i,j} \mathbf{SCO}(\mathbf{T}_{ai}, \mathbf{T}_{bj})$

3.3 Operation Example

Figure 1–3 show operation examples. Figure 1 and 2 are examples of Japanese-English translation. In Figure 1, with only statistical data on the target language text, 'court' may be chosen as an equivalent translation of *"kooto"* because the pair of 'court–judge' co-occurs most frequently in the target language. However with DMAX Criteria, the equivalent translation of *"kooto"* is selected correctly.

- The expression which co-occurs with *"kooto"* most frequently in the source language is *"nekutai"*.
- The pair of the equivalent translation of *"kooto"* and the one of *"nekutai"* which co-occurs most frequently in the target language is 'coat–tie'.
- As a result, *"kooto"* is translated into 'coat'.

It is the same as shown in Figure 2. A pair of 'basket–ball' co-occurs most frequently in the target language. But using DMAX Criteria, giving attention first to the most frequent pair in the source language text, *"kotori–kago"* can gain the correct equivalent translation 'cage'. Next, a pair of *"mizu–booru"* decides 'bowl' as an equivalent translation of *"booru"*. Finally, correct translation can be acquired in this way.

Figure 3 shows the translation process and the statistics for co-occurrence of another English-Japanese translation example.

ENG:	`The ceiling of the court was cleaned quite well.`		
		⇓	
JAP:	*Saibansho–no*		*tenjoo–wa*
	court	of	ceiling subj. marker
	kireini		*souji–sareteita.*
	quite well		be cleaned

In this case, the English words 'court' and 'clean' have two meanings respectively.

'court'

saibansho *a room or building in which law cases can be heard and judged*

kooto *(a part of) an area specially prepared and marked for various ball games, such as tennis*

'clean'

souji–suru *to clean rooms*

kuriiningu–suru *to clean clothes with chemicals instead of water*

A pair of *"kooto–kuriiningu"* co-occurs most frequently in the target language, so the sentence may be translated into *"Kooto–no tenjoo-ha kireini kuriiningu-sareteita."*. But using DMAX Criteria, 'ceiling' is selected as a basis for the equivalent translation selection of 'court', and *"saibansho"* is selected as an equivalent translation of 'court' by the comparison between statistics for co-occurrence on the pairs of *"tenjoo–saibansho"* and *"tenjoo - kooto"*.

4 Calculation of Linguistic Statistics for Semantic Interpretation

In language understanding systems or machine translation systems through semantic expressions, one suitable meaning must be selected out of the ones described in a dictionary according to an entry word. However in conventional systems the meaning selection mechanism isn't robust and cannot select the most suitable meaning only by logical restrictions described in the dictionaries. We presented a new method for the equivalent translation selection in the former chapter using statistical data on source language and target language through bilingual dictionary. To apply this method to meaning selection, it is necessary to calculate statistical data on the pairs of each meaning in advance, but there is no means of calculating them automatically.

We have already developed an interlingua-based machine translation system whose interlingua named PIVOT doesn't depend on any particular natural language [Muraki 86, Ichiyama 89, Okumura 91]. In its dictionary, as illustrated in Figure 4., expressions in the source language are mapped onto some interlingua vocabularies (CONCEPTUAL-PRIMITIVE:CP), which are next mapped onto some equivalent translations. Then we propose a new method of computing linguistic statistics for occurrence of meanings automatically using this format of dictionary.

Suppose linguistic statistics on the pairs of expressions in both source and target language texts have already been calculated. In case of translation, when an expression $\mathbf{S}_i$ occurs in the source language text, an equivalent translation $\mathbf{T}_{ijk}$ is decided through the passage of $\mathbf{S}_i \Rightarrow \mathbf{C}_{ij} \Rightarrow \mathbf{T}_{ijk}$, and as a result, $\mathbf{CPC}_{ij}$ is also selected from the CPs corresponding to the expression $\mathbf{S}_i$. Therefore, the linguistic statistics on the pairs of CPs or meanings is nothing but coupling linguistic statistics on the pairs of corresponding expressions in the target language text. Thus, the linguistic statistics Ω on the pairs of the meaning expressions in the dictionary can be obtained as the sum of the linguistic statistics ω on the pairs of target language expressions according to the following equation.

$$\Omega(\mathbf{C}_{am}, \mathbf{C}_{bn}) = \sum_{p,q} \omega_{pq}(\mathbf{T}_{amp}, \mathbf{T}_{bnq})$$

This linguistic statistics can be added to the dictionary in advance, and we can select the meaning in the same way as equivalent translation selection.

5 Conclusion

We proposed a new method DMAX Criteria (Double Maximize Criteria based on Dual Corpora) in this paper. It can select a suitable equivalent translation or meaning using the statistical data extracted from both source and target language corpora even when linguistic restrictions described in the dictionary or grammar cannot. The dependency relation in the source language is reflected in the translated text through bilingual dictionary. Moreover, the method has the following features:

1. It utilizes linguistic statistics as context information in addition to logical restrictions effective for ambiguity resolution.

2. The source of the linguistic statistics is the dual corpora of source and target languages, not the bilingual corpora (the target language text doesn't have to be the translation of the source language text).

3. The linguistic statistics can be computed semi-automatically in advance.

4. The linguistic statistics on the pairs of meaning expressions are computed from the linguistic statistics in source and target language texts with the interlingua-based bilingual dictionary to resolve ambiguity in meaning interpretation.

Based on this method, we have carried out an experiment on a limited-scale translation system, and confirmed effectiveness of the method. We are preparing further experiments on a large-scale dual corpora with PIVOT interlingua dictionary. Their result will be reported on another paper.

6 Acknowledgments

The authors wish to thank Mr. Masao WATARI for his continuous encouragement. The authors also thank the members of Media Technology Laboratory for their good suggestions.

References

[Brown 90] P.F.Brown et al. "A Statistical Approach to Machine Translation", *Computational Linguistics* , Vol.16, No.2, 1990

[Doi 92] S.Doi and K.Muraki "Robust Translation and Meaning Interpretation Mechanism based on Examples in Dictionary", *Proc. of 44th Annual Conference of IPSJ* , 1P-2, 1992 (in Japanese)

[Ichiyama 89] S.Ichiyama "Multi-lingual Machine Translation System," *Office Equipment and Products*, 18-131, pp.46-48, August 1989

[Muraki 86] K.Muraki "VENUS: Two-phase Machine Translation System," *Future Generations Computer Systems*, 2, pp.117-119, 1986

[Muraki 91] K.Muraki and S.Doi "Translation Ambiguity Resolution by using Text Corpora of Source and Target Languages", *Proc. of 5th Annual Conference of JSAI* , 11-7, 1991 (in Japanese)

[Nagao 84] M.Nagao "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle", *Artificial and Human Intelligence, ed. A.Elithorn and R.Banerji, North-Holland* , 1984

[Nomiyama 91] H.Nomiyama "Lexical Selection Mechanism Using Target Language Knowledge and Its Learning Ability", *IPSJ-WG* , NL86-8 , 1991 (in Japanese)

[Okumura 91] A.Okumura, K.Muraki and S.Akamine "Multi-lingual Sentence Generation from the PIVOT interlingua," *Proc. of MT SUMMIT III*, pp.67-71, July 1991

[Sato 90] S.Sato and M.Nagao "Toward Memory-based Translation", *COLING-90* , 1990

[Sumita 90] E.Sumita, H.Iida and H.Kohyama "Example-based Approach in Machine Translation", *InfoJapan'90* , 1990

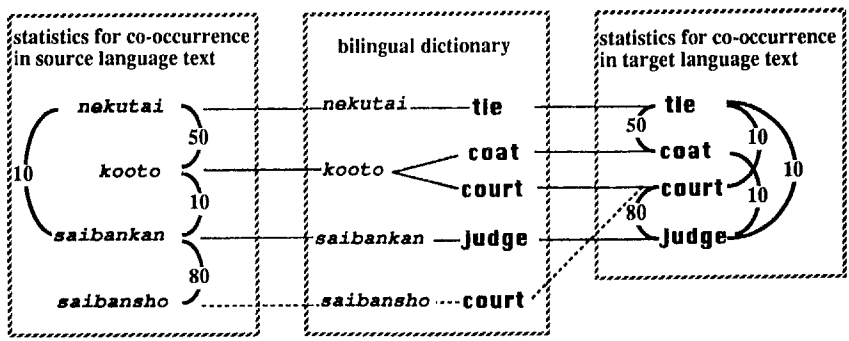


Figure 1 "*Sono saibankan-wa kooto-to nekutai-o katta.*"
'The judge bought a coat and a tie.'

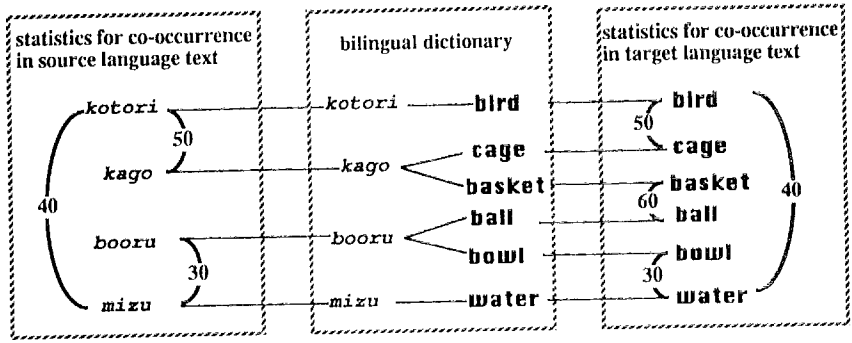


Figure 2 "*Kotori-no kago-ni mizu-o ireta booru-o oita.*"
'I put a bowl filled with water in the bird cage.'

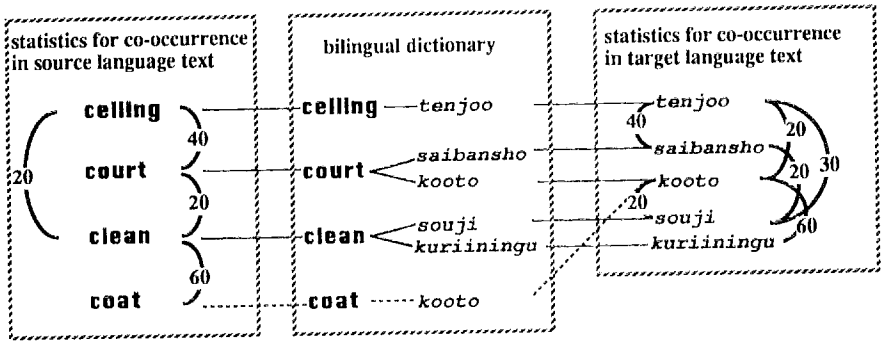


Figure 3 'The ceiling of the court was cleaned quite well.'
"*Saibansho no tenjoo-wa kireini souji-sareteita.*"

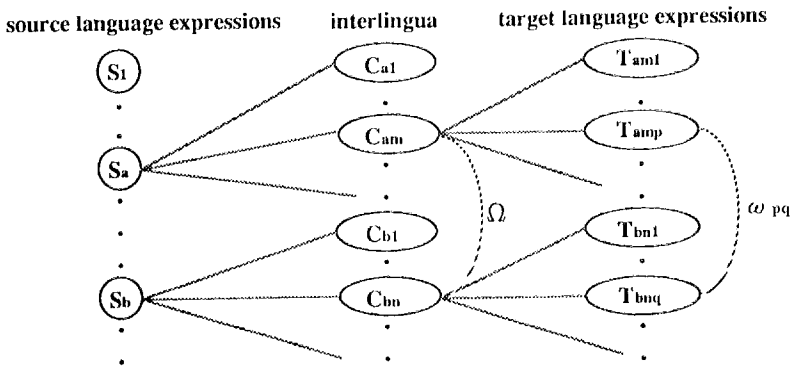


Figure 4 Bilingual dictionary of the interlingua-based translation system