A strategy for solving translation relevant ambiguities in a multi-lingual machine translation system.

Poul Andersen and Annelise Bech

EUROTRA-DK
Njalsgade 80,
DK-2300 Kbh. S
Denmark.

1. Introduction.

Eurotra is a research and development project in machine translation sponsored by the European Commission and the EEC member states. The project was launched in 1984, and its aim is to stimulate research in computational linguistics in Europe, and to produce a running prototype for a multilingual machine translation system towards 1990. This prototype will translate between any two of the nine official languages of the Communities within the subject field of information technology and have a dictionary of approximately 20.000 entries per language.

Most of what we shall say has been inspired by our work as Eurotra researchers, however, the views presented in this paper do not necessarily all reflect the official Eurotra position.

Acknowledgement.
We are indebted to those of our colleagues who have been investigating transfer problems for stimulating and thought-provoking papers on the topic.

A strategy for solving translation relevant ambiguities in a multilingual
machine translation system
Poul Andersen, Annelise Bech
Proceedings of NODALIDA 1987, pages 187-203

2. The Translation System.

Eurotra is designed as a transfer based system. There are separate monolingual components for analysis and generation, and transfer components to link these. This means that we have monolingual components for Danish, Dutch, English, French, German, Greek, Italian, Portuguese, and Spanish and 72 transfer components to link these nine.

In analysis, the task of the monolingual component is to produce an abstract representation of the text. This we call an interface object because this representational object constitutes the input to the transfer component to either one of the other monolingual components. The target language generates a text on the basis of the output from the transfer component.

There are two important principles in the Eurotra design: compositionality and simple transfer. The translation process is compositional, i.e. the translation of a text is a function of the translation of its parts. Simple transfer basically means that the structure of the source language interface object is transferred unchanged to the target component, and that only lexical units change. Ideally, this should result in transfer components that only contain rules specifying the translation of source and target language lexical units, e.g.

```
know -> wissen
know -> kennen
```

for the translation of this English verb into German.

3. Disambiguation in bilingual MT-systems and in multi-lingual MT-systems.

In this paper, we shall present our ideas about how to develop a strategy for solving translation relevant lexical ambiguities in a multi-lingual machine translation system. Here it should be noted that in normal usage, a lexical unit is ambiguous if it has more than one denotation. In our usage, ambiguity is defined contrastively, that is a lexical unit is ambiguous if it has more than one translation into some other language. This was the case in the example already given for the translation of the English verb 'know' into either the German 'kennen' or 'wissen'.

For several reasons, an appropriate strategy for solving translation relevant lexical ambiguities in a multi-lingual machine translation system differs from that which may be adopted in a bilingual system. In a bilingual translation system, the semantic and syntactic similarities of and differences between the two languages can to some degree be accounted for by tuning the source and the target language grammars towards each other. Since the translation relevant ambiguities will be known, a high proportion of the disambiguation needed can be catered for in the source language component by entering a large number of specific readings for each lexical unit in the monolingual dictionary.

In a large multi-lingual system such as Eurotra where the same source language analysis result, i.e. the interface object, constitutes the input to eight different target languages, such a strategy has little attraction. Tuning the monolingual components towards each other would mean that the system would loose in extensibility not only with respect to extension of the grammars of the languages already part of the system, but also with respect to inclusion of new languages into it.

To sum up what has been said so far:

- Ambiguity is defined contrastively, in relation to another language.

- Analysis components should be developed monolingually and consequently such ambiguities cannot be taken into account.

- Transfer components should be kept as simple as possible.

That leaves the burden of disambiguation to the target language generation. As we shall see, this is not in conflict with the claim that generation components also should be developed monolingually. Actually, ambiguity arises bilingually, but can to a large extent be solved monolingually.

4. A strategy for disambiguation.

We propose a strategy where the basic principles are:

1) Disambiguation in analysis is restricted to disambiguation based on morphological criteria.

2) Disambiguation in transfer is restricted to those cases where we need access to information from the source language.

3) As the general principle, disambiguation is left to generation.

4.1. Disambiguation in analysis.

Disambiguation based on morphological criteria means that homographs belonging to different word classes, homograph nouns with different genders, and homographs from the same word class but with different inflection patterns are separated out into separate dictionary entries. This distinction automatically follows from the monolingual description necessary for morphological and syntactical analysis.

This means that we get 3 entries for 'like':

I **like** fish	- VERB
I never saw the **like**	- NOUN
People **like** you and me	- CONJUNCTION

Any other distinction is

- arbitrary
- not needed for monolingual description

4.2. Disambiguation in transfer.

Only in relatively few cases do we need access to information from the source language, and most cases can be handled just as well without access to such information.

One example where this information is needed is the translation of 'put' into German or Danish. The English verb is neutral as to horizontal or vertical position, whereas German and Danish have to make a choice between two verbs, 'stellen'/'stille' for vertical position, 'legen'/'lægge' for horizontal position. It is true that you also have the choice of a position-neutral verb like 'anbringen'/'anbringe' with a different stylistic value, corresponding to English 'place', but let us leave that out for the sake of the argument.

Now, if you have the German translations

'sie _________ die Flasche auf den Tisch'
'sie _________ das Buch auf den Tisch'

and you have to choose the right verb, you may in both sentences use 'stellen' as well as 'legen'. Only, bottles are normally placed in a vertical position on a table and books in a horizontal position, so if nothing was specified in the English text, you would choose the translations

'sie stellte die Flasche auf den Tisch'
'sie legte das Buch auf den Tisch'

Only if the English text had specified e.g. 'she laid the bottle on the table' or 'she stood the book on the table', would you choose the other possibilities, i.e.

'sie legte die Flasche auf den Tisch'
'sie stellte das Buch auf den Tisch'

Incidentally, this example is very dependent on the context. If the item is placed on a shelf, what is normal changes - books are <u>normally</u> put in a vertical position, whereas bottles are put in a horizontal position, at least in a wine cellar. So,

'she put the bottle on the shelf' (= 'on the rack')

translates into

'sie legte die Flasche in das Regal'

and

'she put the book on the shelf'

translates into

'sie stellte das Buch in das Regal'

If we could solve this ambiguity during generation, we would just need two simple rules for English -> German

put -> stellen
put -> legen

and correspondingly for English -> Danish

put -> stille
put -> lægge

and then leave it to generation to rule out the wrong translation. But we need the information that the source language had a neutral verb, and we also need information about the kind of object and about the place of location.

At present, we do not know how to distinguish between words like 'book' and words like 'bottle' nor how to distinguish between locations like 'on the table' and locations like 'on the shelf' in a systematic way. We shall probably have to write rather clumsy translation rules such as

```
put / _, obj | bottle... |, location | table... | -> stellen

put / _, obj | book... |, location | table... | -> legen

put / _, obj | bottle..|, location | shelf, rack..| -> legen

put / _, obj | book... |, location | shelf... | -> stellen
```

which should be read:
'put' translates into 'stellen', if 'put' is followed by an object which is a member of the set mentioned, and a location which contains a noun from the set mentioned.

These 4 rules should be regarded as exception rules to be tried first. If they do not apply, because the object is neither 'book' nor 'bottle', 2 simple rules will apply:

put -> stellen

put -> legen

and we shall get 2 translations of

'he put the newspaper on the table'

1 - 'er stellte die Zeitung auf den Tisch'

2 - 'er legte die Zeitung auf den Tisch'

Of these 2, the first one can be ruled out without having access to the source text, because newspapers not only normally are placed in a horizontal position, they always are - within our linguistic universe.

4.3. Disambiguation in generation.

As the general principle, disambiguation is left to generation. In one respect this is uneconomic because it means that we make more than one translation of ambiguous expressions, only to subsequently rule out the wrong one or the wrong ones. It would be more economic only to make the right translation, of course.

However, in another respect it <u>is</u> economic because in most cases a given ambiguity exists only in relation to some of the other 8 languages making up the system, and in these cases we can benefit from the similarity between the languages when there is no ambiguity.

If, for example, we want to translate the English verb 'adopt' into German, Danish and French, we have at least 3 translations into German and Danish:

1 - They **adopted** a child
┌ Sie **adoptierten** ein Kind
└ De **adopterede** et barn

2 - He has **adopted** a new method
┌ Er hat eine neue Methode **eingeführt**
└ Han har **indført** en ny metode

3 - The Council **adopted** the proposal
┌ Der Rat **verabschiedete** den Vorschlag
└ Rådet **vedtog** forslaget

But into French we can use the same translation of the verb in all 3 cases:

1 - Ils ont adopté un enfant
2 - Il a adopté une nouvelle méthode
3 - Le Conseil a adopté la proposition

If we disambiguate in analysis, we get 3 entries in the English dictionary, adopt_1, adopt_2 and adopt_3. We then need 3 rules from English to French:

```
adopt_1 -> adopter
adopt_2 -> adopter
adopt_3 -> adopter
```

The French might have drawn the same distinction, and we would get:

```
adopt_1 -> adopter_1
adopt_2 -> adopter_2
adopt_3 -> adopter_3
```

If, however, we do not carry disambiguation this far in analysis, we can manage with only one rule from English to French:

```
adopt -> adopter
```

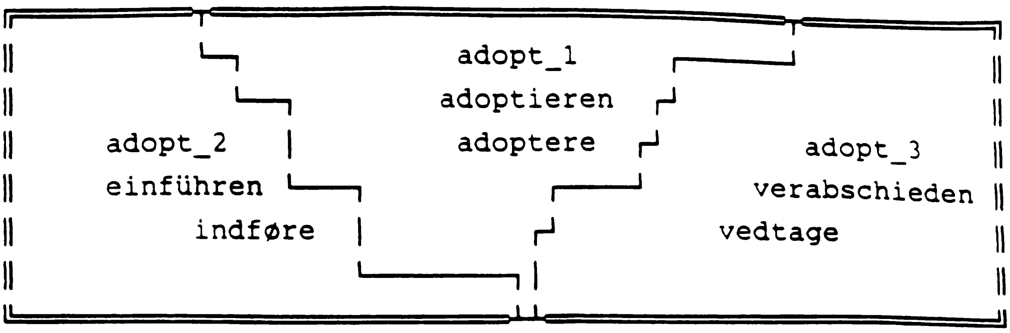
But would it not be convenient to have separate entries 'adopt_1', 'adopt_2' and 'adopt_3' for translating into German and Danish? -

```
        ┌ adoptieren
adopt_1 |
        └ adoptere

        ┌ einführen
adopt_2 |
        └ indføre

        ┌ verabschieden
adopt_3 |
        └ vedtage
```

This would work if 'adopt' always translates into 'einführen' and 'indføre', or into 'verabschieden' and 'vedtage', respectively, that is if the lexical structure of Danish and German were the same:



However, this is not the case, and furthermore the example is too simplified and more translations are needed than just three. In accordance with the principle of simple transfer, we prefer to leave the problem to generation and just write simple, context-free transfer rules:

```
English -> German : adopt -> adoptieren
                    adopt -> einführen
                    adopt -> verabschieden

English -> Danish : adopt -> adoptere
                    adopt -> indføre
                    adopt -> vedtage
```

So, we leave the problem to generation. Monolingually in the target language, we are presented with a choice of three different verbs:

```
    ┌ adoptierten ----┐
Sie ┼ führten --------┼ ein Kind  -- ein
    └ verabschiedeten ┘

                             ┌ adoptiert
Er hat eine neue Methode ┼ eingeführt
                             └ verabgeschiedet

        ┌ adoptierte ----┐
Der Rat ┼ führte --------┼ den Vorschlag  -- ein
        └ verabschiedete ┘
```

and we must make a choice without having access to source language information.

In our example, the necessary rules may be formulated in the dictionary entries in the monolingual German dictionary:

```
(lu=adoptieren, sem_feat_object=+human,-adult)
(lu=einführen, sem_feat_object=+abstract v +concrete,-human)
(lu=verabschieden, sem_feat_object=+admin v +human,+adult)

(lu=Kind, sem_feat=+human,-adult)
(lu=Knabe, sem_feat=+human,-adult,+masculin)
(lu=Mädchen, sem_feat=+human,-adult,-masculin)
(lu=Methode, sem_feat=+abstract)
(lu=Vorschlag, sem_feat=+admin)
(lu=Beamter, sem_feat=+human,+adult)
```

'lu' is short for 'lexical unit'. This approach is based on a marking of all nouns with semantic features so that the selection of a verb can be made dependent on the semantic features of its arguments, i.e. its subject, direct object or indirect object.

The assignment of semantic features may create some problems. For instance, 'verabschieden' is not only a translation of the English verb 'adopt', but also of 'dismiss':

dismiss -> verabschieden

e.g.

'The manager dismissed the official'

translates into

'Der Direktor verabschiedete den Beamten'

However, this is also catered for by assigning two possible semantic feature sets of the object: either '+admin' or '+human,+adult'.

Developing a multi-lingual MT-system is a very delicate task. As has already been pointed out, it is important to have some very clear principles that are motivated and consistent, and that will hold not only for a small prototype system but also allow for extension in terms of lexical coverage and in terms of inclusion of new languages. Yet at the same time, various pragmatic considerations are also necessary.

The implicit principle in the above discussion has been that disambiguation is carried out in generation and based on the semantic features of the context. However, what if we happen to have two linguistic expressions, two lexical units, following each other and both are ambiguous when translated into some language? Then the disambiguation of the first may depend on the semantic features of the second, and the disambiguation of the second may depend on the semantic features of the first. This might create an infinite loop.

Here we are helped by a compositional and context-free translation strategy, however. First all the parts of a sentence are translated, only then do we look at the various combinations. Sometimes this may create problems, but such problems are due to 'true' ambiguities, i.e. ambiguity in the normal usage of the term, that could not have been solved anyway. Suppose for example that we have the following rules from English into German:

```
discard -> verwerfen
discard -> verabschieden
master -> Lehrer
master -> Original (i.e. master copy)
```

and the following German dictionary entries:

```
(lu=verwerfen, semfeat_object= -animate)
(lu=verabschieden, semfeat_object=+human)
(lu=Lehrer, sem_feat=+human)
(lu=Original, sem_feat= -animate)
```

The English sentence:

She discarded the master

will in the first place, compositionally, get four translations:

Sie verwarf den Lehrer
Sie verwarf das Original
Sie verabschiedete den Lehrer
Sie verabschiedete das Original

Of these, two will be ruled out because there is no match between the semantic features of the verb and the object, and two will survive:

Sie verwarf das Original
Sie verabschiedete den Lehrer

The English sentence actually has these two meanings so we should get two translations. However, only one of these gives the intended meaning, but to find this, the system has to look beyond the sentence or to draw on information about text-type just as a human translator would. We shall not eleborate on that here.

In general though, we can rely on nouns being less ambiguous than verbs. This means that in practice we can to a large extent rely on the semantic features of nouns when disambiguating verbs. In the 'adopt' example above, there are no big problems in translating 'child', 'proposal', and 'method' into German, Danish, and French.

So far we have been concerned with contextually determined ambiguities. Within these, we may distinguish between

1. ambiguities that depend on the semantic context

and

2. ambiguities that depend on the syntactic context.

We have seen some examples of the first type and now we shall turn to the second.

The translation of the English verb 'know' into German, French, and Danish is dependent on whether its object is a clause or a noun phrase. Yet also here we can have context free translation rules:

```
English -> German:
     know -> kennen
     know -> wissen

English -> French:
     know -> connaitre
     know -> savoir

English -> Danish:
     know -> kende
     know -> vide
```

and monolingual dictionary entries:

```
German:
     (lu=kennen, object=np)
     (lu=wissen, object=clause)

French:
     (lu=connaitre, object=np)
     (lu=savoir, object=clause)

Danish:
     (lu=kende, object=np)
     (lu=vide, object=clause)
```

which will yield the correct translations:

```
                 ┌ Ich kenne die Frau
I know the woman ├ Je connais la femme
                 └ Jeg kender kvinden
```

┌ Ich weiss, dass sie schn ist
I know that she is beautiful ├ Je sais qu'elle est belle
└ Jeg ved, at hun er smuk

Here again, disambiguation in analysis would not really help us, as we would not get a one-to-one correspondence.
'know + clause' translates into 'savoir', but
'know + NP' may also translate into 'savoir', as in

I know my lesson -> Je sais ma lecon

Neither does the correspondence between 'savoir' and 'wissen'/'vide' hold here, as German and Danish use a modal verb 'knnen'/'kunne':

┌ Ich kann meine Lektion
I know my lesson |
└ Jeg kan mine lektier

The translation of 'know' also demonstrates the need for a proper analysis of the text to be translated including resolution of pronoun references, as the criterion for the choice between 'kennen'/'wissen', 'connaitre'/'savoir' and 'kende'/'vide' is the structure of the antecedent of a pronoun, e.g.

┌ Sie ist schn, das weiss ich
She is beautiful, I know it ├ Elle est belle, je le sais
└ Hun er smuk, det ved jeg

┌ Sie hat ein Problem und ich kenne es
She has a problem and I know it ├ Elle a un probl
me et je le connais
└ Hun har et problem og jeg kender det

Apart from contextually determined ambiguities, we also have inherent ambiguities. This distinction should be seen as an operating distinction in an MT-system. It might be argued that there is an inherent semantic difference between 'adopt' in the sense 'adopt a child' and in 'adopt a proposal', but this is not really of much relevance so long as 'adopt' in the 'proposal'-sense can never take 'child' as an object, nor can 'adopt' in the 'child'-sense take 'proposal' as an object.

Operationally, we want to defer as much as possible to contextually determined ambiguities, as these are better controlled and more interesting from an MT point of view. What is left as inherent semantic ambiguities are consequently those cases where a word has more than one translation in the same context. Generally speaking, contextually determined ambiguities become inherent semantic ambiguities when the context is not informative enough. In these cases, disambiguation typically may be based on information about texttype.

E.g. the English noun 'pipe' translates into Danish 'fløjte', 'pibe' and 'rør'. In the following sentences, the context can be used for disambiguation:

She played the pipe -> Hun spillede på fløjte
She smoked a pipe -> Hun røg pibe
The pipe leaked -> Røret var utæt

However, a sentence like

8 pipes had been ordered

is translated into three equally correct sentences:

Der var blevet bestilt 8 fløjter
Der var blevet bestilt 8 piber
Der var blevet bestilt 8 rør

We must produce only one translation, and only one of the three translations actually convey the intended meaning. In cases like this we would have to apply a lexical preference mechanism, stating that in our text-type - information technology - the last translation is most likely to be the correct one. This mechanism might be based on the following text-type and dictionary information:

```
text-type=information technology > sem_feat=technology,...
text-type=arts  > sem_feat=literature, music, ...

(lu=fløjte, sem_feat=music)
(lu=rør, sem_feat=technology)
```

5. Final remarks.

To conclude, we sum up the principles of our strategy for solving lexical ambiguities in a multi-lingual machine translation system where we want to have the analysis and generation components developed monolingually and to keep the transfer components as simple as possible:

- Lexical disambiguation performed in the source language component is minimalistic in the sense that it is restricted to dealing only with morphologically based ambiguities, i.e. cases of homography where we can distinguish between separate lexical units on the basis of wordclass, gender, and/or inflectional pattern.

- Lexical disambiguation in transfer is restricted to those cases where the target language needs access to semantic information embedded in the source language lexical unit which is not recoverable to the target language on the basis of semantic and/or syntactic context.

- The rest of the disambiguation is to be resolved in target language generation.

From the point of view of efficiency, it might be claimed that a less restrictive approach to disambiguation in transfer would be preferable. Resolving more ambiguities in transfer means that as few translations as possible of a source language lexical unit are input to the target component, and the analysis and generation components can still be developed monolingually. However, such a strategy implies a vast increase in the size and the complexity of the transfer components - the number of which will always be much greater than that of monolingual components in a multilingual system. Therefore, having the target language disambiguate according to the strategy we have outlined here appears to us to be the soundest approach. As we have argued and exemplified, a large number of different types of lexical ambiguity problems lends themselves to being resolved in the course of target language generation in accordance with the principle of truly monolingually based language components.