Poul Andersen

How Close Can We Get to the Ideal of Simple Transfer in Multi-lingual Machine Translation (MT)?

Abstract

The ideal of simple transfer aims at restricting transfer rules to the exchange of unstructured lexical entities—the terminal leaves in the tree structure that is output from monolingual analysis. All information that is not lexicalised in the source language is represented as features to be transferred unchanged to the target language. In EUROTRA this ideal is approached through a centrally coordinated research within various phenomena which are supposed to be of translational relevance, i.e. having language specific surface manifestations. The outcome of this research ideally is to agree on a uniform treatment of these phenomena across languages, thus leading to simple transfer.

The paper makes a non-exhaustive overview over problems solved, problems under investigation, known but outstanding problems, and on this basis introduces a discussion of what will remain as unsolvable problems within an essentially sentence-based MT-system.

1 Introduction

MT-systems traditionally are classified into transfer-based systems and interlingual systems, as illustrated by figure 1 and figure 2, resp., on the next pages. The Interface Structure (IS) for some language is an annotated tree structure, where information is encoded as structure + features:

```
structure                  features

    A                      B = {attribute 1 = value X,
____|___                        attribute 2 = value Y,
|      |                        ...                 }
B    __C__
     |   |
     D   E
```

Proceedings of NODALIDA 1989, pages 103-113

```
                        ---------
                       |   L2    |
                       |  text   |
                       |_________|
                          |     /\
                analysis  |     | synthesis (= generation)
                       ___V_____|___
                      |     L2      |
        transfer      |  interface  |  transfer
        L1 -> L2      |  structure  |  L3 -> L2
        |------------>|_____________|<-----------|
        |   transfer   |           |  transfer   |
        |   L2 -> L1   |           |  L2 -> L3   |
        |  |-----------|           |----------|  |
    ____|__V_____                        _____V__|____
   |      L1       |  transfer L1 -> L3  |     L3      |
   |   interface   | ------------------> |  interface  |
   |   structure   |  transfer L3 -> L1  |  structure  |
   |_______________| <------------------ |_____________|
         /\    |                           /\    |
analysis |     | synthesis       analysis  |     | synthesis
        _|____V_                          _|____V_
        | L1    |                         | L3    |
        | text  |                         | text  |
        |_______|                         |_______|
```

Figure 1: Schematic representation of transfer-based multi-lingual MT

```
                         -----------
                        |    L2     |
                        |   text    |
                        |___________|
                           |     /\
                  analysis |     | synthesis
                    _______V_____|______
                   |   interlingual     |
    analysis       |  representation    |  analysis
   |-------------->|____________________|<-------------|
   |      synthesis    |             | synthesis       |
   |    |--------------|             |------------|    |
  _|____V_                                       _V____|_
 | L1    |                                      | L3    |
 | text  |                                      | text  |
 |_______|                                      |_______|
```

Figure 2: Schematic representation of interlingual multi-lingual MT

EUROTRA is conceived as a transfer-based system, which may seem less appropriate for an MT system comprising 9 languages in all combinations, thus leading to the construction of 72 transfer modules on top of 9 analysis and 9 synthesis modules, instead of just having one interlingua, 9 analysis and 9 synthesis modules.

What we want to show, is that the distinction between transfer- and interlingua-based systems should not be pushed too hard, especially if an interlingua is not perceived as a natural language-like representation but as any kind of information encoding that is neutral with respect to a source language and a target language.

The ideal in transfer is sometimes described as simple lexical transfer, which means that the lexical values are the only information in the interface structure that is not shared by source and target language and which consequently has to be changed by a transfer component, whereas all other information is represented language-independently in an interlingua. Actually, the greater part of lexical transfer may also be dispensed with through the inclusion of a comprehensive terminological component that is treated interlingually.

As the IS representation may be split up into structure information and feature information, we shall treat these independently and distinguish between

1. Transfer of structure

and

2. Transfer of features

2 Transfer of Structure

Here we distinguish between three possibilities:

2.1 Simple transfer = interlingua (= no explicit transfer)

```
L1       A                      L2       A'
     ____|____                       ____|____
     |       |       <=====>         |       |
     B     __C__                     B'    __C'__
           |   |                           |    |
           D   E                           D'   E'
```

2.2 Deletion/insertion of node

```
L1       A                      L2       A'
     ____|____                       ____|____
     |       |       <=====>         |       |
     B     __C__                     B'      E'
           |   |
           D   E
```

2.3 Reordering of elements

```
L1      A                     L2      A'
    ----|----                     ----|----
    |       |      <=====>        |       |
    B     __C__                 __C'__    B'
          |   |                 |    |
          D   E                 D'   E'

L1      A                     L2      A'
    ----|----                     ----|----
    |       |      <=====>        |       |
    B     __C__                   B'    __C'__
          |   |                         |    |
          D   E                         E'   D'
```

2.1 Simple Transfer

This is the unproblematic case where there is isomorphy between source language and target language or where this isomorphy is achieved between output from source language analysis and input to target language synthesis. How this isomorphy is achieved, is described in 2.2 and 2.3 below.

2.2 Deletion/Insertion of Node

2.2.1 Direct/Indirect Government

We have to delete, respectively insert, a node in cases where we have direct government by a verb of a noun phrase in one language corresponding to indirect government through a preposition in another language, e.g.

```
              VP                             VP
        ______|______   <======>    _______|________
        |           |               |              |
        V           NP              V              PP
                    |                          ____|____
                    N                          |       |
                                               P       NP
                                                       |
                                                       N

EN : (He)  trusted    her     DA : (Han) stolede  på   hende
DA : (Han) betragtede hende   EN : (He)  looked   at   her
```

The solution is to featurise all valency bound prepositions, without regard to whether they have a correspondence or not in one or more other languages, and delete the preposition and the PP-node from the IS representation:

```
        VP                                 VP
   -----|----                         -----|-----
   |        |     Danish analysis     |         |
   V        PP    --------------->    V         NP {pform='på'}
   |    ----|----                     |         |
'stole' |       | Danish synthesis 'stole'      N
        P       NP <---------------             |
        |       |                             'pige'
      'på'      N
                |
              'pige'
```

This featurisation of the preposition is accompanied by a feature in the IS dictionary entry for the verb:

```
{da_lu = 'stole', valency = subject_object, pform_of_object = 'på'}
```

It should be noted that it is not always without problems to distinguish between valency bound complements, where the preposition is deleted from the structure, and free modifiers, which at present keep their preposition.

Another related problem is indirect government by a verb through an NP, which is described in detail in Susanne Nøhr Pedersen's paper 'The Treatment of Support Verbs and Predicative Nouns in Danish'.

2.2.2 Function Words vs. Inflectional Endings

Another example of deletion/insertion of nodes are function words in one language which correspond to inflectional endings in other languages, e.g. articles with nouns and auxiliaries with verbs. Here again the problem is solved by representing the information contained in the function word as a feature on the content word or its projection, i.e. the NP or the VP.

A problem arises e.g. in country names, which take the definite article in French but go without article in Danish, English and German. In these cases we would prefer to block the automatic transfer of definiteness and leave it to the target language to calculate its surface representation. Modified country names, again, might have their definiteness transferred, as in 'a united Europe' or 'das Europa der Nachkriegszeit', although this is not without problems. Determination and quantification in general is a very complicated subject to be treated contrastively, and at pręsent it is being investigated as a special research topic within EUROTRA.

2.2.3 Featurisation vs. Structural Representation

There are strong advantages in representing as much information as possible as features and thus reducing complex structural transfer,—an approach I personally favour. In EUROTRA there is, however, a certain opposition against removing too much from the structure. The argumentation is that most surface words can be modified, and it is more convenient to represent a modification of a node in a structure than to modify information that has been featurised. As two examples of surface expressions that might be featurised—and actually were featurised, but now must be present as nodes in the IS representation—we may mention modal verbs and demonstrative pronouns.

2.3 Reordering of Elements

2.3.1 Reordering at NP Level

Reordering of elements occurs at NP level, where a modifier may precede the noun or follow after the noun, and where the ordering in different languages also differ according to the category of the modifier:

```
            Adjective + Noun     <=====>    Noun + Adjective

     DA : den  blå    himmel
     EN : the  blue   sky          FR : le ciel  bleu
     DE : der  blaue  Himmel
     but
     FR : la   petite fille

          NP modifier + Noun      <======>     Noun   +   NP/PP modifier

                                   EN : the inhabitants of the country
     DA : landets   indbyggere     DE : die Einwohner      des Landes
                                   FR : les habitants   du      pays
```

The solution is to have a common, language-independent ordering (referred to as 'canonical ordering') of the elements in the IS representation, and do the necessary reordering in analysis and synthesis:

```
Canonical ordering
---------------------------|--------------------------
|                          |                         |
governor           complements (valency-bound)    modifiers (free)

Canonical ordering of NPs

        NP                            NP
   -----|-----                   -----|-----
   |         |                   |         |
   N         AP                  N         NP {case=genitive}
             |                             |
             ADJ                           N {case=genitive}
```

```
+ featurisation of prepositions:

          NP                                 NP
     _____|_____  EN/FR analysis        _____|_____
     |         |  ==============>       |         |
     N        PP                        N        NP {case=genitive}
           ___|___  EN/FR synthesis               |
           |     |  <==============               N {case=genitive}
           P    NP
           |     |
      EN : of    N
      FR : de
```

2.4 Reordering at Sentence Level

Two examples of reordering at sentence level:

```
sentence 1

      NP + Vaux + Vmain + NP             NP + Vaux + NP    + Vmain

DA :  Hun   har   spist   brødet
EN :  She   has   eaten   the bread  DE : Sie  hat das Brot gegessen
FR :  Elle  a     mangé   le pain

sentence 2

      AdvP   +   Vaux +  NP     +    Vmain  +  PP
DA :  I går      blev    forslaget  vedtaget  af Rådet

      AdvP   +   Vaux +  NP          +    PP    +  Vmain
DE :  Gestern    wurde   der Vorschlag  vom Rat   verabschiedet

      AdvP   +   NP            +    Vaux + Vmain +  PP
EN :  Yesterday  the proposal      was     adopted  by the Council
FR :  Hier       la  proposition   a été   adoptée  par le Conseil
```

In analysis, articles and auxiliary verbs are featurised and removed from the structure, and the fact that the sentence is in passive voice is marked as a feature at the top node. At present, we do not use a refined set of semantic case roles but restrict ourselves to a numbering of arguments, where i.a. the subject of a sentence in active voice is labelled 'argl' and the object is labelled 'arg2'. The maximum number of arguments in a sentence is 4.

Somewhat simplified, and without feature information, the IS representation of the two sentences looks like this:

```
                                   S
   ----------------------------|-----------------------------
   |              |                        |                    |
governor      argument 1               argument 2           modifier
   |     (= 'semantic' subject)  (= 'semantic' object)          |
   |              |                        |                    |
```

sentence 1

DA : spise	hun	brød	
DE : essen	sie	Brot	
EN : eat	she	bread	
FR : manger	elle	pain	

sentence 2

DA : vedtage	Rådet	forslag	i går
DE : verabschieden	der Rat	Vorschlag	gestern
EN : adopt	the Council	proposal	yesterday
FR : adopter	le Conseil	proposition	hier

The canonical ordering of the elements is in itself fairly straightforward and poses no major problems. What creates problems may be differences between languages and differences between language groups in analysis of some constituent, e.g. as complement or modifier. This is the reason why we are very wary of introducing a too ambitious approach in assigning case roles, as this would give rise to inconsistencies between assigment carried out in different language groups.

3 Transfer of Features

Here again we distinguish between three possibilities:

3.1 Features which are transferred unchanged.

3.2 Features which are not transferred but calculated again in the target language or found in the target dictionary.

3.3 Features with an explicit translation in the transfer component.

A feature has the form {attribute=value}, and what is transferred unchanged, calculated or translated explicitly is only the value of the feature.

In the first two cases no explicit transfer is needed, which means that we have simple transfer or interlingual treatment. In 2.2.2 above we mentioned definiteness as an example of a surface phenomenon that gives rise to both the first two types of transfer of features. In many cases morpho-syntactic definiteness may express semantic definiteness in a consistent way across languages, and in these cases we may transfer the value for the 'definiteness' attribute unchanged. Some (sub)categories, however, allow only of one of the paradigmatic set of values for definiteness, and this value may not be the same for different languages. In these cases the value is not transferred but found in the target dictionary — or in the target grammar if it is possible to generalise over a class of words, cf. the example mentioned in 2.2.2 with country names, e.g. ‘la France’ {definiteness=definite} versus ‘Frankrig’ {definiteness=absent}.

In general, feature values which are not transferred, are typically bound to a lexical value, e.g. gender and semantic features on nouns and pforms on verbs (i.e. the preposition used in a valency bound PP), as well as other valency frame information, including restrictions on the semantic features of valency bound complements. These values are looked up in the dictionary. The value for ‘gender’ is then used to generate the correct form of modifying adjectives and determiners, and the valency information in the dictionary entry for a verb is matched with the available information on the complements. It is also used to connect the complement by means of the correct preposition in the target language.

Where there is more than one translation of a verb, the valency information in the target dictionary is used to decide which translation matches the structure of the IS object, which may be transferred unchanged, e.g.

```
      EN IS-object     transfer EN->DA    DA IS-object
                       ==============>
            S                                  S
--------- |---------                  --------- |--------
|          |         |                |          |        |
governor  subject   object          governor  subject  object
|          |         |                |          |        |
know       NP      Scompl            vide       NP     Scompl
               ----|-----------                    -----|----------
               |       |      |                    |       |       |
              verb   subject etc.                 verb   subject  etc.
```

```
lexical transfer rules

{en_lu = know}  => {da_lu = kende}
{en_lu = know}  => {da_lu = vide}

Danish target dictionary

{da_lu = kende, da_isframe = np_subject_np_object}
{da_lu = vide, da_isframe = np_subject_scompl_object}
```

We do not need to transfer the valency information of ‘know’, as only ‘vide’ matches the transferred IS-object, due to the restriction on ‘da_isframe’ in the Danish target dictionary (the description here is somewhat simplified).

We want to restrict explicit translation in the transfer component to lexical values in the narrow sense as uninflected wordforms. But this lexical transfer may also be reduced through an interlingual approach to certain categories of words. We have already mentioned function words such as noun determiners and auxiliary verbs, which are featurised and given an interlingual description.

But where we really hope to save a lot of explicit transfer rules is in the treatment of terms. The implementation of terminology is just being started, but we hope to treat the greater part of the planned 20.000 entries dictionaries for each of the 9 EUROTRA languages as terms without entering them in the 72 transfer dictionaries. Terms are coded centrally with their valency frames and they are assigned a unique term-number to be used as reference (instead of the lexical value) by all language groups. The general frame description specifies how many, and which, valency bound arguments a given term takes, and each language group then complements this with language-specific information about which preposition, surface case etc. is used together with which argument.

4 Conclusion

As conclusion, we shall show by means of an example how far towards an interlingual interface structure we in principle have advanced. The example is somewhat simplified, and the actual state of implementation in EUROTRA may differ slightly from this presentation of an 'ideal' implementation.

```
French text
Hier,  la France  a adopté  la proposition  du Conseil.

Danish translation
I går  vedtog     Frankrig  Rådets          forslag.
```

Please note that the translation involves i.a. —

- reordering of 'la France/Frankrig' and 'a adopté/vedtog'
- reordering of 'la proposition/forslag' and 'du Conseil/Rådets'
- change of morpho-syntactic tense/aspect from 'parfait simple' to 'imperfektum', — both expressing the same semantic past
- change of morpho-syntactic definiteness of 'la France/Frankrig'
- change of the surface manifestation of the definiteness of 'du Conseil/Rådet' and 'la proposition/forslag' (in the latter case only being expressed through a preposed genitive)

```
IS representation

                                S
 -------------------------------|-----------------------------
 |                  |                  |                     |
governor          subject            object               modifier
{fr_lu=adopter/   {term=184}     _______|______          {fr_lu=hier/
 da_lu=vedtage,                  |             |          da_lu=i_går,
 time=past}                   governor      modifier      position=
                              {term=200,    {term=237}    initial}
                               number=
                               singular,
                               definiteness=
                               definite}
```

The only difference in the IS representation for the two languages are the lexical values for the two non-terms, so the only explicit transfer rules needed are the following:

```
FR-DA transfer dictionary

{fr_lu = adopter}  ==>  {da_lu = vedtage}
{fr_lu = hier}     ==>  {da_lu = i_går}
```

All other information is contained in the two monolingual dictionaries:

```
FR dictionary                     DA dictionary

{term = 184,                      {term = 184,
 fr_lu = 'France',                 da_lu = 'Frankrig',
 fr_definiteness = definite,       da_definiteness = absent,
 fr_gender = feminin,              da_gender = neuter,
 fr_number = singular}             da_number = singular}

{term = 200,                      {term = 200,
 fr_lu = 'proposition',            da_lu = 'forslag',
 fr_gender = feminin}              da_gender = neuter}

{term = 237,                      {term = 237,
 fr_lu = 'Conseil',                da_lu = 'Rådet',
 fr_definiteness = definite,       da_definiteness = definite,
 fr_gender = masculin,             da_gender = neuter,
 fr_number = singular}             da_number = singular}

{fr_lu = 'adopter',.              {da_lu = 'vedtage',
 fr_isframe = subject_object}      da_isframe = subject_object}

{fr_lu = 'hier'}                  {da_lu = 'i_går'}
```

For clarity of exposition, only information relevant to our example is included here.

In this example we distinguish between 'definiteness' and 'fr_definiteness'/ 'da_definiteness'. The idea is that a feature may have a language-independent attribute name in cases when it expresses semantic information to be carried over, and the same attribute name with a language prefix in cases when the value is not semantically significant but concerns monolingual wellformedness. The distinction between universal features and monolingual features is currently made in EUROTRA by means of uniform attribute names + prefixes, which enables/disables matching, but this way of using the same attribute name with or without prefix is not implemented.

EUROTRA-DK
Njalsgade 80
DK-2300 København S.
poul@eurotra.dk