Sorting the chaff from the wheat in corpus-based machine translation

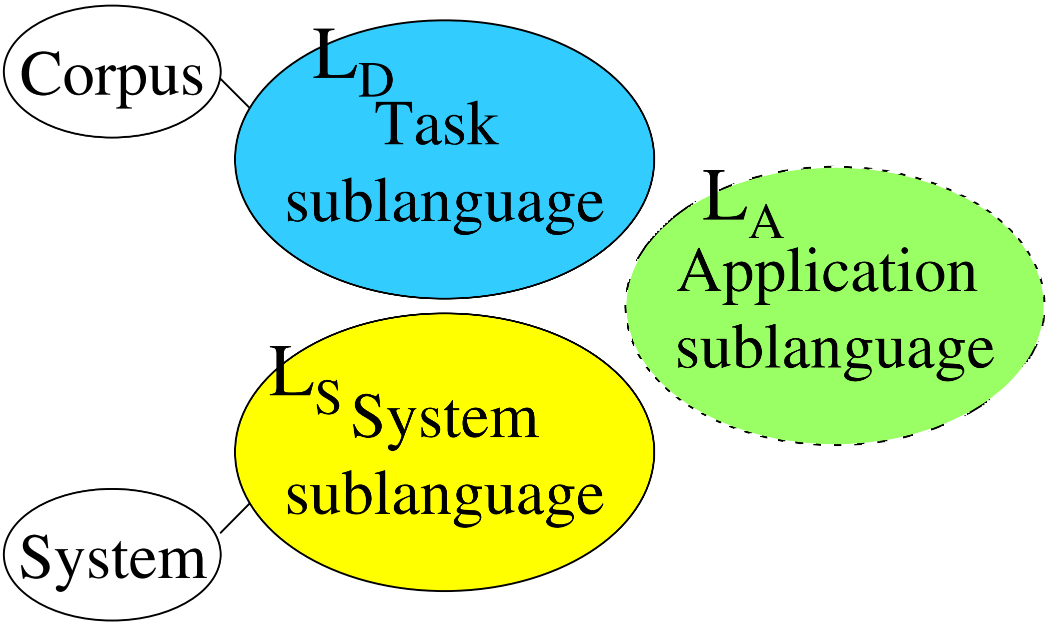
Lars Ahrenberg

NLPLab

Department of Computer and Information Science, Linköpings universitet

Corpus-based machine translation

- Adaption (or creation) of a MT system to a (restricted) translation task on the basis of a translation corpus that represents the task well, resulting in an application system.



Question

What is the relation of the application sublanguage to the given task language?

Application sublanguage

- The application sublanguage L_A in general cannot coincide with the task sublanguage L_D
  - For dissemination applications, restrict the source side of L_D to achieve a better agreement with L_S translations (i.e. the controlled language approach);
  - For assimilation applications, restrict the target side of L_A to achieve a better coverage of L_D inputs.

  Note: This paper is only concerned with the second approach!
- Relation to task and system sublanguages
  - source side coincides with the source side of the task sublanguage as far as possible
  - target side must be within the bounds of an extended version of the system sublanguage
- Quality criteria
  - Intelligibility,
  - Accuracy,
  - Well-formedness

The Problem

- Compared with human translators, MT systems are generally
  - less knowledgeable,
  - less flexible,
  - more error-prone;
- In corpus-based MT, human translations are used as training data where free translations are not uncommon; thus, there is a high probability that the system is trained on data that it cannot reproduce;
- Moreover, automatic procedures that are used to populate the system's databases will produce erroneous data.

The Solution

- Extendible core system providing a L_S of fair contrastive structural coverage;
- Methods for sorting data;
- Methods and guidelines for normalising data.

An example: T4F

T4F is a direct translation system satisfying an Alignment principle that requires source sentence and translation to be aligned at the lexical level by 1-n alignments ($n \geq 0$). Ordering rules at the target side may introduce discontinuous alignments.

Translation is performed in five stages:

1. Tokenisation (= identification of units for lexical lookup)
2. Tagging
   - part-of-speech, with morphological subcategorisation,
   - dependency function and dependency argument,
   - contextual features,
3. Transfer (= lexical lookup, using tags as filters)
4. Transposition (= reordering) and Filtering (= removal of translation options based on constraint rules)
5. Probabilistic ranking of surviving options

Transfer dictionaries are read off an alignment file. Alignments are derived automatically, using a supervised initial alignment produced interactively.

Some illustrative T4F results (BLEU scores)

System configuration	Training	Test
T4F ATIS ia1	0.977	0.759
T4F ATIS ia2	0.981	0.781
T4F Access XP aa	0.530	0.307
T4F Access XP ia	0.683	-
Bl Access XP aa	0.524	0.295
Bl Access XP ia	0.592	-

Bl, the baseline, does not use the tagging, filtering and transposition modules.
aa = interactive + automatic alignment,
ia = interactive alignment only

References:

http_//www.ida.liu.se/~nlplab/koma/publications.shtml
Maria Holmqvist (2003). Översättningssystemet T4F – en implementation för ATIS-domänen.
Lars Ahrenberg & Maria Holmqvist (2004). Back to the future? The case for English-Swedish Direct Machine Translation. Paper presented at the workshop on Recent Advances in Scandinavian Machine Translation (RASMAT 2004), Uppsala, 23. April, 2004.

Comments on the scores:

- The ATIS corpus is a case of close translation (in fact, machine translation) while the Access XP corpus is a case of human translation with frequent translation shifts. This explains most of the differences in scores;
- Accurate alignment is important for the Baseline and absolutely critical for T4F. With erroneous alignments at the syntactic level, the difference between T4F and the Baseline almost disappears.

Relation of application sublanguage to task and system:

$$L_D = \{<ds_i, dt_i>; ds_i \in SL_D, dt_i \in TL_D, \mathrm{Transl}_D(ds_i,dt_i) \}$$

$$L_S = \{<ss_i, st_i>; ss_i \in SL, dt_i \in TL, MT_S(ss_i,st_i) \}$$

$$L_A = \{<as_i, at_i>; as_i \in SL_D, at_i \in \mathrm{Ext}(TL_S), \mathrm{Transl}_D(as_i,at_i), MT_S (as_i,at_i) \}$$

$\mathrm{Transl}_D(s,t)$ means that sentence t is an accurate translation of sentence s in the context of the given task;

MT_S (s,t) means that t is a sentence that a specific MT system can produce as output, given s as input.

Defining sublanguages

- The system sublanguage should include a fair coverage of common SL constructions and their TL translations based on contrastive knowledge of SL and TL.
- Selection criteria for core translations:
  - **Semantic stability**: selected translations should be near equivalents to the source item across a large set of contexts;
  - **Structural equivalence**: selected translations should not generally require complex structural rearrangements;
  - **Low rank**: include translations for high rank constructions only if lower rank translations do not suffice;
  - **Distinguishable use conditions**: the contexts that favour the use of one translation, should be possible to differentiate from the contexts that favour its alternatives, if not descriptively, at least in practice;

cont.

- **Avoid mutual translation dependencies**: Translations should be independent in the sense that their conditions of use do not refer mutually to other translations. By generally allowing the two translations below, the head part of the NP cannot be translated independently of the genitive attribute, and vive versa:

  English: the contents | *of the file*
  Swedish 1: filens innehåll
  Swedish 2: innehållet i filen

  Note: mutual dependencies at one rank require the recognition of units at a higher rank.

- Additional selection criteria for the application system
  - **Frequency:** Selected translations should preferably be among the most common ones in use for the given task;
  - **Stylistic appropriateness:** Selected translations must be appropriate for the task language.

Sorting the chaff from the wheat

- Normalising test data
  - Test data should be representative of the application. This means that the data used for testing should not be taken at random from the corpus, but normalised according to the quality criteria of the application, and the system's capabilities;
- Normalising training data
  - Ideally, training data should be normalised too, but this is hard to achieve, since normalisation requires human labour. However, corpus data that have been found to contain ”dangerous” features can be normalised instead of just being thrown away.
- Sorting training data
  - Added material and paraphrases are especially difficult for automatic alignment. Moreover, we wish to weed out analysis errors introduced by the tagging module, as these introduce inaccurate syntactic correspondences.

Normalising translations

E: A CSS merely allows you to specify the formating of each XML element without much control over the output.

S: Med en CSS-formatmall kan du bara ange formateringen av varje XML-element men inte styra resultatet i särskilt hög grad.

S1: En CSS-formatmall tillåter bara att du anger formateringen av varje XML-element utan särskilt mycket kontroll över resultatet.

S2: En CSS tillåter dig bara att ange formateringen av varje XML-element utan mycket kontroll över resultatet.

E: It displays individual charts for Buchanan and Davolio.

S: Enskilda diagram för Berg och Danielsson visas.

S1: Det visas individuella diagram för Berg och Danielsson.

S2: Den|det visar individuella diagram för Buchanan och Davolio.

Sorting data

Length differences often signal additions, or paraphrases

E: XML tags are case-sensitive .
S: XML-märken gör skillnad mellan versaler och gemener .

E: white space can be used throughout the document to enhance readability .
S: det går att använda blanksteg i hela dokumentet för att göra det mer lättläst .

E: the default is to exclude hidden data from totals .
S: standardinställningen är att dolda data exkluderas , men den inställningen kan du ändra .

E: you can select two or more custom groups to create a higher-level grouping .
S: du kan markera två eller flera egna grupper om du vill skapa grupper på högre nivå .

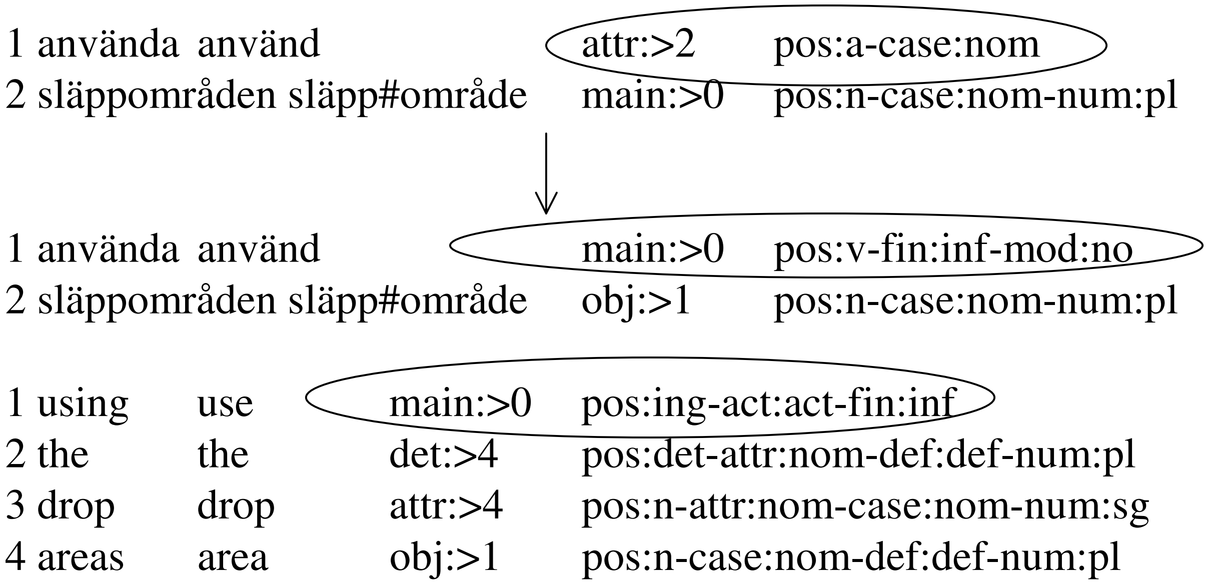
E: you can also remove a lower-level custom group .
S: du kan också ta bort en anpassad grupp på låg nivå .

Multi-word alignments either signal (i) paraphrastic translations to be weeded out (or normalised), or (ii) multi-word lexicalised phrases, to be included in the system dictionaries.

1, <From this standpoint, Med ett sådant perspektiv>
8, <in the name of, med hänvisning till>
10, <is no exception, Så är fallet även i>
25, <on the issue of, i fråga om>
39, <big for their boots, styva i korken>
54, <are in confrontation, som står emot varandra>
95, <As far as I know, enligt min bedömning>
129, <on the other hand, å andra sidan>
133, <a host of things, allt och lite till>
161, <that have a financial impact, viktiga beslut av>
164, <good luck to him, jag önskar honom all framgång>
179, <That just goes to show, Där ser ni>
223, <Ladies and gentlemen, Mina damer och herrar>
225, <will be put to the vote, omröstningen kommer att äga rum>
236, <have your cake and eat it, få allt på samma gång>
250, <main objective is, att man framför allt>
255, <Now we come to, Jag vill ta upp>
257, <The way China develops, Utvecklingen i Kina>
285, <In the first place, För det första>
324, <the information I have, vad jag kan säga>
325, <You have the floor, Ordet är ert>
337, <now has the floor, Ordet går till>
354, <This is why, Det är anledningen till att>

Data from Philipp Koehn's Europarl corpus. Extraction from an I*Link alignment file.

Erroneous analyses may yield inaccurate tag correspondences and should be weeded out



```
1 använda  använd               attr:>2     pos:a-case:nom
2 släppområden släpp#område     main:>0     pos:n-case:nom-num:pl
                              ↓
1 använda  använd               main:>0     pos:v-fin:inf-mod:no
2 släppområden släpp#område     obj:>1      pos:n-case:nom-num:pl

1 using     use         main:>0     pos:ing-act:act-fin:inf
2 the       the         det:>4      pos:det-attr:nom-def:def-num:pl
3 drop      drop        attr:>4     pos:n-attr:nom-case:nom-num:sg
4 areas     area        obj:>1      pos:n-case:nom-def:def-num:pl
```

Such erroneous analyses can be detected by comparing the correspondences resulting from alignment with a store of standard correspondences associated with the core system.