Automatic Induction of Transfer Rules using Aligned Bilingual Corpora

Yvette Graham

National Center for Language Technology
School of Computing
Dublin City University

May 31 2006
BiText Project

Project Aims
- Automatically generate a number of aligned bilingual corpora (bi-texts)
- Acquire existing bilingual corpora.
- Automatically annotate the bi-texts with LFG c-structures and f-structures.
- Automatically induce transfer rules from the corpora.

Traditionally, transfer rules for machine translation have been hand-coded - takes time and resources.
Automatically inducing such transfer rules from large bilingual corpora could provide an efficient way of acquiring such transfer rules.
SL sentence parsed and f-structure achieved.
Transfer Rule database is searched for a matching rule
  ▶ The SL sentence f-structure must match the LHS of the transfer rule.
  ▶ If multiple rules match, then the most specific and most likely one is selected.

Recursively apply transfer rules to any parts of the f-structure not taken care of by the initial rule.
Perhaps adding a default rule, for cases in which there is no matching transfer rule for a given f-structure.
Adding this information to the RHS of the initial rule produces the TL f-structure.
This f-structure is then used to generate the TL sentence.
Some advantages of this approach to MT

The f-structure notation allows us to infer desirable transfer rules for f-structure relations that correspond to non-adjacent words and thus would not be discovered in the string-based world.

For example:

- **Es gibt ...**
  - There is ...
  - (lit. *It gives ...*)

- **Es scheint ... zu geben.**
  - There seems [...] to be.
  - (lit. *It seems ... to give.*)

  - - Riezler, Maxwell (2006)

Transfer rules allow for mapping between different linguistic types.

For example, extraction of the phrase pair:

- **zutiefst dankbar ⇒ a deep appreciation**

Filtering of f-structure phrases based on consistency of linguistic types would find this mapping invalid because of the incompatibility of types A and N for adjective *dankbar* and nominal *appreciation*.

  - - Riezler, Maxwell (2006)
Automatically Inducing Transfer Rules

- F-structure annotated bilingual corpora are used to extract transfer rules for a given language pair.
- Relevant sentences and their translations are extracted from such corpora and analysed.
- The similarities between the SL f-structures are examined ⇒ generalisation about the structure of *the sentence to be translated* and form the LHS of the transfer rule.
- The similarities amongst the TL f-structures are examined ⇒ a generalisation about the structure of *the translation of the sentence* and forms the RHS of the rule.
- This results in a transfer rule meaning: given a SL sentence whose f-structure matches the LHS of the rule, the f-structure of its translation is likely to be that of the RHS of the rule.
Example: Automatically Inducing Transfer Rule for the verb “talk” from English to French

- Search the English side of all available English/French bi-text corpora for sentences in which the main predicate is talk.
- Retrieve both the English sentences and their matching French sentences.
- Examine the f-structures of the sentences looking for regularities amongst the English f-structures and then similarities amongst the f-structures of their translations.
- For Example: the search of the corpora could return the following three sentence pairs
  1. The man talks to the dog.  *L’homme parle au chien.*
  2. He talks to Mary all day.  *Il parle à Marie toute la journée.*
  3. The president talks to the nurse.  *Le président parle à l’infirmière.*
The man talks to the dog.

L’homme parle au chien.

F-Structures for Sentence Pair 1
F-Structures as Equations: Sentence Pair 1

**English F-Structure:**

```
{ SUBJ (0, 1)  
  PRED (0, ‘talk(SUBJ, OBL OBJ)’)  
  OBL (0 ,2)  
  PRED (1, ‘man’)  
  DEF (1, +)  
  NUM (1, sg)  
  PRED (2, ‘to(OBJ)’)  
  OBJ (2, 3)  
  PRED (3, ‘dog’)  
  NUM (3, SG)  
  DEF (3, +) }
```

**French F-Structure:**

```
{ SUBJ (0, 1)  
  PRED (0, ‘parler(SUBJ, OBJ)’)  
  OBJ (0 ,2)  
  PRED (1, ‘homme’)  
  DEF (1, +)  
  NUM (1, sg)  
  GEND (1, masc)  
  PRED (2, ‘chien’)  
  DEF (2, +)  
  NUM (2, sg)  
  GEND (2, MASC)  
  PFORM (2, à_) }
```
Linking the Content of the Two F-Structures

- Examine the values of the predicates of the English f-structure
- Translate each of the values using a bilingual dictionary or word list
- Search for the translated predicates in the French F-structure
- Record links between the f-structures

1. Predicate values of the English f-structure

\[
\begin{align*}
\text{PRED} (0, 'talk\langle\text{SUBJ, OBL OBJ}\rangle') \\
\text{PRED} (1, 'man') \\
\text{PRED} (2, 'to\langle\text{OBJ}\rangle') \\
\text{PRED} (3, 'dog')
\end{align*}
\]

2. Dictionary Look-Up

- talk $\Rightarrow$ parler
- man $\Rightarrow$ homme
- to $\Rightarrow$ à
- dog $\Rightarrow$ chien

3. Translations in the French f-structure predicate values:

\[
\begin{align*}
\text{PRED} (0, 'parler\langle\text{SUBJ, OBJ}\rangle') \\
\text{PRED} (1, 'homme') \\
\text{PRED} (2, 'chien')
\end{align*}
\]

4. Resulting links between f-structures:

- $0 \Rightarrow 0$ write as variable A
- $1 \Rightarrow 1$ write as variable B
- $3 \Rightarrow 2$ write as variable C
## Linking the Content of the F-Structures

### English F-Structure:

<table>
<thead>
<tr>
<th>Property</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>SUBJ</code></td>
<td>(A, B)</td>
</tr>
<tr>
<td><code>PRED</code></td>
<td>(A, ‘talk⟨SUBJ, OBL OBJ⟩’)</td>
</tr>
<tr>
<td><code>OBL</code></td>
<td>(A ,2)</td>
</tr>
<tr>
<td><code>PRED</code></td>
<td>(B, ‘man’)</td>
</tr>
<tr>
<td><code>DEF</code></td>
<td>(B, +)</td>
</tr>
<tr>
<td><code>NUM</code></td>
<td>(B, sg)</td>
</tr>
<tr>
<td><code>PRED</code></td>
<td>(2, ‘to⟨OBJ⟩’)</td>
</tr>
<tr>
<td><code>OBJ</code></td>
<td>(2, C)</td>
</tr>
<tr>
<td><code>PRED</code></td>
<td>(C, ‘dog’)</td>
</tr>
<tr>
<td><code>NUM</code></td>
<td>(C, SG)</td>
</tr>
<tr>
<td><code>DEF</code></td>
<td>(C, +)</td>
</tr>
</tbody>
</table>

### French F-Structure:

<table>
<thead>
<tr>
<th>Property</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>SUBJ</code></td>
<td>(A, B)</td>
</tr>
<tr>
<td><code>PRED</code></td>
<td>(A, ‘parler⟨SUBJ, OBJ⟩’)</td>
</tr>
<tr>
<td><code>OBJ</code></td>
<td>(A ,C)</td>
</tr>
<tr>
<td><code>PRED</code></td>
<td>(B, ‘homme’)</td>
</tr>
<tr>
<td><code>DEF</code></td>
<td>(B, +)</td>
</tr>
<tr>
<td><code>NUM</code></td>
<td>(B, sg)</td>
</tr>
<tr>
<td><code>GEND</code></td>
<td>(B, masc)</td>
</tr>
<tr>
<td><code>PRED</code></td>
<td>(C, ‘chien’)</td>
</tr>
<tr>
<td><code>DEF</code></td>
<td>(C, +)</td>
</tr>
<tr>
<td><code>NUM</code></td>
<td>(C, sg)</td>
</tr>
<tr>
<td><code>GEND</code></td>
<td>(C, MASC)</td>
</tr>
<tr>
<td><code>PFORM</code></td>
<td>(C, à_)</td>
</tr>
</tbody>
</table>
Examine the left side of the rules

LHS for Sentence Pair 1:

- **SUBJ (A, B)**
- **PRED (A, ‘talk⟨SUBJ, OBL OBJ⟩’)**
- **OBL (A, 2)**
- **PRED (B, ‘man’)**
- **DEF (B, +)**
- **NUM (B, sg)**
- **PRED (2, ‘to⟨OBJ⟩’)**
- **OBJ (2, C)**
- **PRED (C, ‘dog’)**
- **NUM (C, sg)**
- **DEF (C, +)**

LHS for Sentence Pair 2:

- **SUBJ (A, B)**
- **PRED (A, ‘talk⟨SUBJ, OBL OBJ⟩’)**
- **OBL (A, 2)**
- **PRED (B, ‘pro’)**
- **CASE (B, nom)**
- **GEND (B, male)**
- **NUM (B, sg)**
- **PRED (2, ‘to⟨OBJ⟩’)**
- **OBJ (2, C)**
- **PRED (C, ‘Mary’)**
- **SPEC (D, 5)**
- **QUANT (5, E)**
- **PRED (E, ‘all’)**

LHS for Sentence Pair 3:

- **SUBJ (A, B)**
- **PRED (A, ‘talk⟨SUBJ, OBL OBJ⟩’)**
- **OBL (A, 2)**
- **PRED (B, ‘president’)**
- **DEF (B, +)**
- **NUM (B, sg)**
- **PRED (2, ‘to⟨OBJ⟩’)**
- **OBJ (2, C)**
- **PRED (C, ‘nurse’)**
- **NUM (C, sg)**
- **DEF (C, +)**

The common equations among the set of f-structures gives us the LHS of the transfer rule:

- **SUBJ (A, B)**
- **PRED (A, ‘talk⟨SUBJ, OBL OBJ⟩’)**
- **OBL (A, 2)**
- **PRED (2, ‘to⟨OBJ⟩’)**
- **OBJ (2, C)**
Examine the right side of the rules

The common equations among the set of f-structures gives us the RHS of the transfer rule:

\[
\begin{align*}
\text{subj} & (A, B) \\
\text{pred} & (A, \text{`parler}(\text{subj}, \text{obj})`) \\
\text{obj} & (A, C)
\end{align*}
\]
Resulting Transfer Rule

\[
\{ \text{SUBJ} (A, B), \text{PRED} (A, \text{‘talk'}(\text{SUBJ}, \text{obl OBJ})), \text{OBL} (A, 2), \text{PRED} (2, \text{‘to'}(\text{OBJ})), \text{OBJ} (2, C) \} \Rightarrow \{ \text{SUBJ} (A, B), \text{PRED} (A, \text{‘parler'}(\text{SUBJ}, \text{OBJ})), \text{OBJ} (A, C) \}
\]

Finally, replacing any remaining f-structure labels with variables:

\[
\{ \text{SUBJ} (A, B), \text{PRED} (A, \text{‘talk'}(\text{SUBJ}, \text{obl OBJ})), \text{OBL} (A, D), \text{PRED} (D, \text{‘to'}(\text{OBJ})), \text{OBJ} (D, C) \} \Rightarrow \{ \text{SUBJ} (A, B), \text{PRED} (A, \text{‘parler'}(\text{SUBJ}, \text{OBJ})), \text{OBJ} (A, C) \}
\]
Current Project Status

Work to date:
- Compiled a list of available bilingual corpora.
- Developed a small trilingual corpus of sentences in German, English and French.
- Used XLE to parse the sentences producing their f-structures in prolog format.
- Manually compiled word-lists for the example sentences.

Currently working on:
- Implementing and testing different algorithms to extract transfer rules from the f-structures of aligned bilingual sentences for a given language pair.
Some Future Work

Further develop the induction algorithm

Use the Human Centre Corpus to induce transfer rules and as validation for the induction algorithm

- French English (900 sent. approx)

Scale things up further to use a larger f-structure annotated corpus

Use the induced grammars developed by members of the LFG group to annotate existing bi-texts

Use these annotated bitexts to induce transfer rules