Constrained Recombination in an Example-based Machine Translation System

Monica Gavrila

University of Hamburg
Faculty of Mathematics, Informatics and Natural Sciences

EAMT
(Research Paper)
29.05.2011
Contents

1 The Framework

2 MT Systems

3 Experiments

4 Conclusions
Example-based Machine Translation

- Translation by analogy (Nagao, 1984).
- A (small) parallel aligned corpus is enough: database of examples.
- Three steps: matching, alignment and recombination.
- Several Approaches: linear, template-based, hybrid etc.

Template: (...) gave (...) up ⇔ (...) a abandonat (...)

- Languages: Romanian, German, English
- Romanian as under-resourced language
Contents

1. The Framework
2. MT Systems
3. Experiments
4. Conclusions
The Implemented MT Systems

1. \textit{Lin − EBMT}
   - The EBMT baseline system
   - A linear EBMT system

2. \textit{Lin − EBMT^{REC+}}
   - Extends \textit{Lin − EBMT}
   - Hybrid system (linear + template-based)
   - Word-order constraints are used in the recombination step. The constraints are extracted from templates.
Lin – EBMT Matching

- Recursive approach
- Based on surface-forms
- Based on the longest common subsequence (LCS) algorithm (Bergroth et al, 2000)
- A token-index is used to reduce the matching space.
LCS Similarity (LCSS)

Given two strings - $s_1$ and $s_2$ - the LCSS measure is calculated as

$$LCSS(s_1, s_2) = LCSST(s_1, s_2) - P \times noTG,$$

where

$$LCSST(s_1, s_2) = \frac{\text{Length}(LCS(s_1, s_2))}{\text{Length}(s_1)},$$
Example

Input $s_1 = "\text{Saving names and phone numbers ( Add name )}"$
Sentence in the corpus $s_2 = "\text{Erasing names and numbers}"$

$LCS(s_1, s_2) = "\text{names and numbers}"$

$LCSS(s_1, s_2) = \frac{3}{9} - 0.01 \times 1 = 0.323$. 
Lin – *EBMT*: Alignment

- Uses GIZA++ results and the longest TL aligned subsequence are used

**LCS: ”technical regulations standards”**

**Alignments**
- ”*technical - tehnice*” (position 8 in TL),
- ”*regulations - reglementări*” (position 7 in TL) and
- ”*standards - standarde*” (position 23 in TL)

We use further the sequences: ”*reglementări tehnice*” and ”*standarde*”.
Lin – EBMT: Recombination

- Input the "the bag of word sequences" \( \{w_1, w_2, \ldots, w_n\} \) provided by the alignment step.
- The result is the needed translation.
- Uses a "recombination matrix"
The Recombination Matrix

Let \( A = a(i, j) \) be the "recombination matrix". If the outcome of the alignment is \( n \) word-sequences \( \{w_1, w_2, \ldots, w_n\} \) which form the output and are not necessarily different, with \( w_i = w_{i_1}w_{i_2} \ldots w_{i_{last}} \), then \( A \) is a square matrix of order \( n \) that is defined as follows:

\[
A = \begin{cases} 
-3, & \text{if } i = j; \\
-2, & \text{if } i <> j, \\
\frac{2 \cdot \text{count}(w_{i_{last}}w_{j_1})}{\text{count}(w_{i_{last}})+\text{count}(w_{j_1})}, & \text{if } w_{i_{last}}w_{j_1} \text{ is not in the corpus;}
\end{cases}
\]

(3)

else.
$w_i, \ 1 \leq i \leq n, \text{ is a sequence.}$

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>...</th>
<th>$w_i$</th>
<th>...</th>
<th>$w_j$</th>
<th>...</th>
<th>$w_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td></td>
<td>-3</td>
<td>$a(1,2)$</td>
<td>...</td>
<td>$a(1,i)$</td>
<td>...</td>
<td>$a(1,j)$</td>
<td>...</td>
</tr>
<tr>
<td>$w_2$</td>
<td>$a(2,1)$</td>
<td></td>
<td>-3</td>
<td>...</td>
<td>$a(2,i)$</td>
<td>...</td>
<td>$a(2,j)$</td>
<td>...</td>
</tr>
</tbody>
</table>
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ...
| $w_i$ | $a(i,1)$ | $a(i,2)$ | ... | ... | ... | ... | ... | ... | $a(i,n)$ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ...
| $w_j$ | $a(j,1)$ | $a(j,2)$ | ... | $a(j,i)$ | ... | ... | ... | -3 | ... | $a(j,n)$ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ...
| $w_n$ | $a(n,1)$ | $a(n,2)$ | ... | $a(n,i)$ | ... | $a(n,j)$ | ... | ... | ... | -3 |
The Recombination Matrix - 2

\[
\begin{array}{cccccccc}
  & w_1 & w_2 & \ldots & w_i & \ldots & w_j & \ldots & w_n \\
 w_1 & & -3 & a(1,2) & \ldots & a(1,i) & \ldots & a(1,j) & \ldots & a(1,n) \\
w_2 & a(2,1) & & -3 & \ldots & a(2,i) & \ldots & a(2,j) & \ldots & a(2,n) \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
w_i & a(i,1) & a(i,2) & \ldots & \ldots & -3 & \ldots & \color{red}a(i,j) & \ldots & a(i,n) \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
w_j & a(j,1) & a(j,2) & \ldots & a(j,i) & \ldots & \ldots & -3 & \ldots & a(j,n) \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
w_n & a(n,1) & a(n,2) & \ldots & a(n,i) & \ldots & a(n,j) & \ldots & \ldots & -3 \\
\end{array}
\]

\(w_i, \ 1 \leq i \leq n,\) is a sequence.
### The Recombination Matrix - 2

<table>
<thead>
<tr>
<th></th>
<th>w1</th>
<th>w2</th>
<th>...</th>
<th>wi</th>
<th>...</th>
<th>wj</th>
<th>...</th>
<th>wn</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td></td>
<td></td>
<td></td>
<td>a(1,i)</td>
<td></td>
<td>a(1,j)</td>
<td></td>
<td>a(1,n)</td>
</tr>
<tr>
<td>w2</td>
<td>a(2,1)</td>
<td>-3</td>
<td>...</td>
<td>a(2,i)</td>
<td></td>
<td>a(2,j)</td>
<td></td>
<td>a(2,n)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wi</td>
<td>a(i,1)</td>
<td>a(i,2)</td>
<td>...</td>
<td></td>
<td>-3</td>
<td>a(i,j)</td>
<td></td>
<td>a(i,n)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wj</td>
<td>a(j,1)</td>
<td>a(j,2)</td>
<td>...</td>
<td>a(j,i)</td>
<td></td>
<td></td>
<td>-3</td>
<td>a(j,n)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wn</td>
<td>a(n,1)</td>
<td>a(n,2)</td>
<td>...</td>
<td>a(n,i)</td>
<td></td>
<td>a(n,j)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ w_i, \ 1 \leq i \leq n, \text{ is a sequence.} \]
The Framework | MT Systems | Experiments | Conclusions
--- | --- | --- | ---

**The Recombination Matrix - 2**

<table>
<thead>
<tr>
<th></th>
<th>w1</th>
<th>w2</th>
<th>...</th>
<th>wi</th>
<th>...</th>
<th>wj</th>
<th>...</th>
<th>wn</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td>-3</td>
<td>a(1,2)</td>
<td>...</td>
<td>a(1,i)</td>
<td>...</td>
<td>a(1,j)</td>
<td>...</td>
<td>a(1,n)</td>
</tr>
<tr>
<td>w2</td>
<td>a(2,1)</td>
<td>-3</td>
<td>...</td>
<td>a(2,i)</td>
<td>...</td>
<td>a(2,j)</td>
<td>...</td>
<td>a(2,n)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>wiwj</td>
<td>a(j,1)</td>
<td>a(j,2)</td>
<td>...</td>
<td>a(j,i)</td>
<td>...</td>
<td>-3</td>
<td>...</td>
<td>a(j,n)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>wn</td>
<td>a(n,1)</td>
<td>a(n,2)</td>
<td>...</td>
<td>a(n,i)</td>
<td>...</td>
<td>a(n,j)</td>
<td>...</td>
<td>-3</td>
</tr>
</tbody>
</table>

\(w_i, \ 1 \leq i \leq n, \) is a sequence.
$w_i$, $1 \leq i \leq n$, is a sequence.

<table>
<thead>
<tr>
<th></th>
<th>w1</th>
<th>w2</th>
<th>...</th>
<th>wi</th>
<th>...</th>
<th>wj</th>
<th>...</th>
<th>wn</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td></td>
<td>-3</td>
<td>a(1,2)</td>
<td>...</td>
<td>a(1,i)</td>
<td>...</td>
<td>a(1,j)</td>
<td>...</td>
</tr>
<tr>
<td>w2</td>
<td>a(2,1)</td>
<td></td>
<td></td>
<td>...</td>
<td>a(2,i)</td>
<td>...</td>
<td>a(2,j)</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>wi</td>
<td>a(j,1)</td>
<td>a(j,2)</td>
<td>...</td>
<td>a(j,i)</td>
<td>...</td>
<td>...</td>
<td>-3</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>wn</td>
<td>a(n,1)</td>
<td>a(n,2)</td>
<td>...</td>
<td>a(n,i)</td>
<td>...</td>
<td>a(n,j)</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
$w_i, 1 \leq i \leq n$, is a sequence.
The Recombination Matrix - 2

<table>
<thead>
<tr>
<th></th>
<th>w1</th>
<th>w2</th>
<th>...</th>
<th>wiwj</th>
<th>...</th>
<th>wn</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td></td>
<td>-3</td>
<td>a(1,2)</td>
<td>...</td>
<td>a(1,j)</td>
<td>...</td>
</tr>
<tr>
<td>w2</td>
<td>a(2,1)</td>
<td></td>
<td>-3</td>
<td>...</td>
<td>a(2,j)</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>wiwj</td>
<td>a(j,1)</td>
<td>a(j,2)</td>
<td>...</td>
<td>...</td>
<td>-3</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>wn</td>
<td>a(n,1)</td>
<td>a(n,2)</td>
<td>...</td>
<td>a(n,j)</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

$w_i$, $1 \leq i \leq n$, is a sequence.
Motivation: use the information which is lost in the recombination step of $Lin - EBMT$;

Mixture of linear and template-based approach;

Matching and alignment remain as in $Lin - EBMT$;

Constraints are set on the values from the recombination matrix, by using information extracted from templates.
Template Extraction

\[
(((TF_{SL})*(VAR_{SL}^{*}))*TF_{SL}*)((TF_{SL})*(VAR_{SL}^{*}))* \leftrightarrow

(((TF_{TL})*(VAR_{TL}^{*}))*TF_{TL})
\]
The input

press and hold clear to delete the characters more quickly.

Matched sentence and alignment

pentru a sterge simultan toate caracterele cand scrieti un mesaj, apasati optiuni si selectati stergeti textul.

to delete all the characters at once when writing a message press options and select clear text.
Template-Example

The input

press and hold clear to delete the characters more quickly.

Template

```
to&&1&& delete&&2&& VAR3 the&&4&& characters&&5&& VAR6 NOALIGN7 VAR8_18 .&&19&& ↔ pentru&&1&& a&&1&& sterge&&2&& VAR6 VAR3 caracterele&&5&& VAR8_18 .&&19&&
```
Constraints

1. **The First-Word-Constraint (C.1):** A constraint C.1 refers to the first word of the output.

2. **TLSide-Template-Constraint (C.2):** The C.2 constraints are deduced only from the TL side of each of the templates extracted.

3. **Whole-Template-Constraint (C.3):** The C.3 constraints are extracted considering each of the templates, together with the input sentence, and the alignment information.

The result: a set $C = \{(\text{word}_i, \text{word}_j)\}$ of constraints: The sequence word$_i$word$_j$ is not allowed.
C.1 Constraints

The input

to delete the characters more quickly press and hold clear.

Template

to&&1&& delete&&2&& VAR3 the&&4&& characters&&5&& VAR6 NOALIGN7 VAR8_18 .&&19&& ↔ pentru&&1&& a&&1&& sterge&&2&& VAR6 VAR3 caracterele&&5&& VAR8_18 .&&19&&
C.2 Constraints

Template

to&&1&& delete&&2&& VAR3 the&&4&& characters&&5&& VAR6 NOALIGN7 VAR8_18 .&&19&&
↔ pentru&&1&& a&&1&& sterge&&2&& VAR6 VAR3 caracterele&&5&& VAR8_18 .&&19&&
New Recombination Matrix

\[ A = \begin{cases} 
-3, & \text{if } i = j; \\
-2, & \text{if } i <> j, \\
\frac{2 \ast \text{count}(w_{ilast} w_{j1})}{\text{count}(w_{ilast}) + \text{count}(w_{j1})}, & \text{if } w_{ilast} w_{j1} \text{ is not in the corpus or } (w_{ilast} w_{j1}) \in C; \\
\text{else.} & 
\end{cases} \] (4)
Another Recombination Matrix

\[ A = \begin{cases} 
-3, & \text{if } i = j; \\
-1, & \text{if } i <> j, \\
-2, & \text{if } w_{i_{last}} w_{j_1} \text{ is not in the corpus; } \\
\frac{2 \cdot \text{count}(w_{i_{last}} w_{j_1})}{\text{count}(w_{i_{last}})+\text{count}(w_{j_1})}, & \text{else.}
\end{cases} \]
Contents

1 The Framework

2 MT Systems

3 Experiments

4 Conclusions
The Experimental Settings

- 2 EBMT systems: $Lin - EBMT$, $Lin - EBMT^{REC+}$
- 2 language pairs, both directions of translations: English-Romanian, German-Romanian
- 1 corpus: RoGER
The Corpus: RoGER

- Developed between 2005 and 2006, at the University of Hamburg, NatS, together with Natalia Eliţa
- Romanian, German, English, Russian
- Manual of an electronic device
- 2333 sentences, between 25K and 27K words
- Manually verified
- No diacritics, some data replaced with meta-notations
Experimental Setting

- Training: 2200 sentences, approx 27 K items, 13 words
  the average sentence length
- Test: 133 sentences, approx 1.6 K items, 12.3 words
  the average sentence length
The Framework | MT Systems | Experiments | Conclusions

BLEU (Papineni et al., 2002) Scores

![Bar graph showing BLEU scores for different systems and language pairs]

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Lin-EBMT</th>
<th>C.1</th>
<th>C.2</th>
<th>C.3</th>
<th>C.1+2</th>
<th>C.1+3</th>
<th>C.2+3</th>
<th>C.1+2+3</th>
<th>C.1+2+3 1:2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romanian-German</td>
<td>0.2867</td>
<td>0.2842</td>
<td>0.2857</td>
<td>0.2891</td>
<td>0.2836</td>
<td>0.2891</td>
<td>0.2875</td>
<td>0.2875</td>
<td>0.2894</td>
</tr>
<tr>
<td>German-Romanian</td>
<td>0.2643</td>
<td>0.2658</td>
<td>0.2682</td>
<td>0.2627</td>
<td>0.2654</td>
<td>0.2627</td>
<td>0.2633</td>
<td>0.2633</td>
<td>0.2646</td>
</tr>
<tr>
<td>Romanian-English</td>
<td>0.3597</td>
<td>0.3695</td>
<td>0.3711</td>
<td>0.3633</td>
<td>0.3712</td>
<td>0.3632</td>
<td>0.3656</td>
<td>0.3656</td>
<td>0.3668</td>
</tr>
<tr>
<td>English-Romanian</td>
<td>0.2997</td>
<td>0.3067</td>
<td>0.3042</td>
<td>0.3083</td>
<td>0.3062</td>
<td>0.3083</td>
<td>0.3073</td>
<td>0.3073</td>
<td>0.3085</td>
</tr>
</tbody>
</table>
Evaluation

Best Score Differences:

- English-Romanian: 0.0088
- Romanian-English: 0.0115
- German-Romanian: 0.0039
- Romanian-German: 0.0027
Contents

1. The Framework
2. MT Systems
3. Experiments
4. Conclusions
Conclusions & Further Work

- Impact of word-order constraints

Further work:

- Additional constraints;
- Priorities for the constraints are used (weighting);
- Different corpus and languages;
- Manual analysis of the data;
- N-grams of several lengths etc.
Thank you for your attention!

Suggestions  ...  Questions  ...