Using Example-Based MT to Support Statistical MT when Translating Homogeneous Data in a Resource-Poor Setting

Sandipan Dandapat, Sara Morrissey, Andy Way, Mikel L. Forcada

CNGL, School of Computing, DCU
Introduction

Over the past two decades, statistical MT (SMT) has shown very promising results

- Requires reasonably good amount of parallel corpora

A large number of languages suffer from the scarcity of large parallel corpora

- Indic languages, Sign languages etc.

Some studies have shown SMT approaches have yielded low translation quality for these poorly resourced languages (Islam et al, 2010; Khalilov et al., 2010).
Introduction

- Domain-specific translation to tackle the issue of scarce resources
  - Very low accuracy within SMT framework for homogeneous domain (Dandapat et. al., 2010)

- Can example-based MT (EBMT) techniques help?
  - EBMT approach can be developed using a limited example base (Somers, 2003)
  - EBMT system works well when training and test data are quite close in nature (Marcu, 2001)
Our Attempt

- We adopt two different EBMT approaches for translating homogeneous data in a resource-poor setting.

I. A compiled approach to EBMT
   - Produces *translation templates* during the training stage (Cicekli and Güvenir, 2001)

II. A novel way of integrating TM into an EBMT system
   - Using a subsentential TM (extracted using an SMT system) in the alignment and recombination stages of an EBMT system
Structure of the Corpus

The size and type of corpora is important for adopting a particular data-driven approach to MT.

We use the IWSLT 2009 English–Turkish corpus to deal with less-resourced homogeneous data.

- The training data is quite small (20k parallel sentences)
- Corpus is comprised of very similar domain-specific sentences

<table>
<thead>
<tr>
<th></th>
<th>Have you ever seen a Japanese movie?</th>
<th>Have you ever tried Japanese food?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (a)</td>
<td>I’d like to see that camera on the shelf.</td>
<td></td>
</tr>
<tr>
<td>2. (a)</td>
<td>I’d like to have it parted on the left.</td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Approach I

- Generalized translation-template-based EBMT
  - **Learning phase:** learn templates from sentence-aligned bitext
  - **Decoding phase:** translate new sentences using the translation templates
Generalized translation-template-based EBMT

**Learning phase** - learns templates from bitext by studying similarities and differences between two example pairs (Cicekli and Güvenir, 2001:p. 58)

*I will drink orange juice → portakal suyu içeceğim  
I will drink coffee → kahve içeceğim*

<table>
<thead>
<tr>
<th>English</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>I will drink</td>
<td>içeceğim</td>
</tr>
<tr>
<td>coffee</td>
<td>kahve</td>
</tr>
<tr>
<td>orange juice</td>
<td>portakal suyu</td>
</tr>
</tbody>
</table>

We assign a probabilistic score \((p)\) to each translation template \(T_i : s_i \rightarrow t_i\)

\[
p(t_i \mid s_i) = \frac{\text{count}(s_i \rightarrow t_i)}{\text{count}(s_i)}
\]

*X ס orange juice → portakal suyu XT 0.33(p)*)
Decoding

∀ untranslated segment in \( s \in \{S\} \)

Template Matching

\( T_i \subset TT \)

Apply \( S_i \rightarrow t_i \) to \( S \)

Translations (partial)

Complete?

\( S = S' \)

Extract top N partial translations

\( S' \)

\( Q \)

\( q \leftarrow q \times p \times w \)

\( \text{numSurfaceWords}(s_i) / \text{length}(s) \)

Ordered set of translations
Approach II

- **EBMT using subsentential TM**
  - **Matching** - finds the closest match with the input sentence
  - **Alignment** - finds translation of the desired segments
  - **Recombination** - combines the translations of the desired segments
Building a Subsentential TM

- We build an auxiliary subsentential TM automatically from the English–Turkish small training corpus
- We use Moses to automatically build this TM
  - Aligned phrase pairs from the Moses phrase table
  - Aligned word pairs based on GIZA++

### Entries in TM from Moses phrase table

<table>
<thead>
<tr>
<th>English</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don't like it</td>
<td>“sevmedim”, “bunu sevmedim”</td>
</tr>
<tr>
<td>I can't sleep well.</td>
<td>“iyi uyuyamıyorum .”</td>
</tr>
</tbody>
</table>

### Entries in TM from word-alignment

<table>
<thead>
<tr>
<th>English</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helps</td>
<td>“vücudun”, “yardım”, “eder”</td>
</tr>
<tr>
<td>Coffees</td>
<td>“kahve”</td>
</tr>
</tbody>
</table>

- We keep all target equivalents sorted according to phrase translation probability
Matching

- We find the closest sentence \( s_c \) from the example base for the input sentence \( s \) to be translated:

\[
 s_c = \arg \max_i \text{score}(s, s_i)
\]

- Edit distance metric to find this closest match sentence:

\[
\text{score}(s, s_i) = 1 - \frac{\text{ED}(s, s_i)}{\max(\|s\|, \|s_i\|)}
\]

\(s: \text{i’d like a present for my mother .}\)

\(s_c: \text{i’d like a shampoo for greasy hair .}\)

- We consider the associated translation \( t_c \) of \( s_c \) to build the skeleton for the translation of the input sentence \( s \):

\(t_c: \text{yağlı saçlar için bir şampuan istiyorum .}\)

GREASY HAIR FOR ONE SHAMPOO I’D-LIKE .
Alignment

- We extract the translation of the non-matching fragments of the input sentence \((s)\).
- To do this, we align three sentences - the input \((s)\), the closest source-side match \((s_c)\) and its target equivalent \((t_c)\).

1. Mark the mismatched portion between input sentence \((s)\) and the closest source-side match \((s_c)\) using edit distance.

\[ s : \text{i’d like a } \text{<present> for } \text{<my mother>} . \]
\[ s_c : \text{i’d like a } \text{<shampoo> for } \text{<greasy hair>} . \]
Alignment

- We extract the translation of the non-matching fragments of the input sentence \( (s) \)
- To do this, we align three sentences - the input \( (s) \), the closest source-side match \( (s_c) \) and its target equivalent \( (t_c) \)

2. We align the mismatched portion of \( s_c \) with its associated translation \( t_c \) using our TM

\[
\begin{align*}
    s : & \text{ i’d like a } \texttt{<present>} \text{ for } \texttt{<my mother>} . \\
    s_c : & \text{ i’d like a } \texttt{<shampoo>} \text{ for } \texttt{<greasy hair>} . \\
    t_c : & \texttt{<1:yağlı saçlar>} \text{ için bir } \texttt{<0:şampuan>} \text{ istiyorum} .
\end{align*}
\]
- The numbers in angle brackets keep track of the order of the appropriate fragments
Recombination

- Substitute, add or delete segments from the input sentence \( s \) with the translation skeleton \( t_c \).

\[
\begin{align*}
s & : \text{i’d like a } <\text{present}> \text{ for } <\text{my mother}>. \\
s_c & : \text{i’d like a } <\text{shampoo}> \text{ for } <\text{greasy hair}>. \\
t_c & : <1:\text{yağlı saçlar}> \text{ için bir } <0:\text{şampuan}> \text{ istiyorum}.
\end{align*}
\]

\[
\begin{align*}
t(\text{my mother}) &= ? \\
t(\text{present}) &= ?
\end{align*}
\]

- We estimate the \( t(\cdot) \) from our subsentential TM.
  - Recursively translating the longest possible matched segment in TM
Experiments

- **Baseline SMT** (using Moses)
- **GEBMT** - baseline experiment with generalized translation template-based EBMT
- **EBMT** - based only on the matching step. Considering closest match target \( t_c \) as the output
- **EBMT\text{\_TM}** - after obtaining the translation skeleton, unmatched segments are translated using subsentential TM
- **English–Turkish data used for experiments**
  - Training Data - 20k sentences (IWSLT’09 training data)
  - Test Data - 414 sentences (IWSLT’09 devset)
Combining the Systems with SMT

- EBMT systems (GEBMT and EBMT\textsuperscript{TM}) sometimes produce correct solutions where SMT fails and vice-versa
- We combine GEBMT and SMT based on the translation score ($q$) for an input test sentence ($s$)
  - If the value of $q$ is greater than some threshold we rely on GEBMT($s$) otherwise we take the output from SMT($s$)
- We call this GEBMT\textsubscript{score} >x + SMT

- We combine EBMT\textsuperscript{TM} and SMT (EBMT\textsuperscript{TM} + SMT) based on two features
  - Fuzzy match score (FMS)
  - The equality in number of mismatched segments in $s$, $s_c$ and $t_c$ (EqUS)
- Rely on EBMT\textsuperscript{TM} output depending on these two features
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Data: 1242 sentences</strong></td>
<td></td>
</tr>
<tr>
<td>SMT</td>
<td>7.63</td>
</tr>
<tr>
<td>GEBMT</td>
<td>6.80</td>
</tr>
<tr>
<td><strong>GEBMT</strong> <em>score&gt;0.3</em> +SMT</td>
<td>7.96</td>
</tr>
<tr>
<td><strong>Training Data: 2184 sentences</strong></td>
<td></td>
</tr>
<tr>
<td>SMT</td>
<td>10.72</td>
</tr>
<tr>
<td>GEBMT</td>
<td>07.21</td>
</tr>
<tr>
<td><strong>GEBMT</strong> <em>score&gt;0.9</em> +SMT</td>
<td>10.83</td>
</tr>
<tr>
<td><strong>GEBMT</strong> <em>score&gt;0.8</em> +SMT</td>
<td>10.99</td>
</tr>
<tr>
<td><strong>GEBMT</strong> <em>score&gt;0.7</em> +SMT</td>
<td>10.76</td>
</tr>
</tbody>
</table>

Accuracy obtained with GEBMT system using very small data

Accuracy obtained with GEBMT system with little more data
# Results

**Accuracy obtained with EBMT<sub>TM</sub> system**

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data: 19,922 sentences</td>
<td></td>
</tr>
<tr>
<td>SMT</td>
<td>23.59</td>
</tr>
<tr>
<td>EBMT</td>
<td>15.60</td>
</tr>
<tr>
<td>EBMT&lt;sub&gt;TM&lt;/sub&gt;</td>
<td>20.08</td>
</tr>
</tbody>
</table>

**System: EBMT<sub>TM</sub> + SMT**

<table>
<thead>
<tr>
<th>Condition</th>
<th>time/percentage EBMT&lt;sub&gt;TM&lt;/sub&gt; used</th>
<th>BLEU(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMS &gt;0.85</td>
<td>35 (8.5%)</td>
<td>24.22</td>
</tr>
<tr>
<td>FMS &gt;0.8</td>
<td>114 (27.5%)</td>
<td>23.99</td>
</tr>
<tr>
<td>FMS &gt;0.7</td>
<td>197 (47.6%)</td>
<td>22.74</td>
</tr>
<tr>
<td><strong>FMS &gt;0.85 &amp; EqUS</strong></td>
<td>24 (5.8%)</td>
<td><strong>24.41</strong></td>
</tr>
<tr>
<td>FMS &gt;0.8 &amp; EqUS</td>
<td>76 (18.4%)</td>
<td>24.19</td>
</tr>
<tr>
<td>FMS &gt;0.7 &amp; EqUS</td>
<td>127 (30.7%)</td>
<td>24.08</td>
</tr>
</tbody>
</table>
Assessment of Error Types

- Incorrect alignment in matching phase
  - Due to erroneous TUs in the subsentential TM
    - $s$: i have a terrible <headache>.
    - $s_c$: i have a terrible <cough>.
    - $t_c$: berbat bir öksürüğüm var.
  
    cough $\rightarrow$ {“öksürük”, “öksürük tedavisi için”} in TM
    - $t'$: berbat bir öksürüğüm var baş ağrısı.

- Incorrect translation produced during decoding
  - Mostly when falling back to word-based translation

- Incorrect morpho-syntactic alignment
  - $s$: do you have a japanese <guidebook>?
  - $s_c$: do you have a japanese <magazine>?
  - $t_c$: japonca bir <0: derginiz> var mı?
  - $t'$: japonca bir rehber kitap var mı?
Observations

- Effect of training data size in EBMT<sub>TM</sub> system
Observations

- GEBMT system has lower accuracy on its own compared to baseline SMT
- Combining GEBMT with SMT has some improvement over SMT
  - relative BLEU improvement of 4.3% with 1242 sentences; less (2.5% relative BLEU) with 2184 sentences

- EBMT\textsuperscript{TM} system has higher score than baseline when the amount of data is small
- With increased data size, SMT performs better compared to EBMT\textsuperscript{TM} system
- Combing EBMT\textsuperscript{TM} and SMT using FMS and EqUS shows improvement over the baseline SMT
Conclusion

- EBMT works better for certain sentences when the amount of available resources is limited.
- Combining EBMT and SMT may be expected to yield a higher score than an individual system.
- Integration of subsentential TM with EBMT improves translation quality.
Future Work

- In order to test the scalability, we plan to use larger training and test data.

- We intend to find more sophisticated features (other than FMS and EqUS) to trigger the use of EBMT system.
Thank You

Questions?