Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited

Miquel Esplà-Gomis  Felipe Sánchez-Martínez
Mikel L. Forcada
{mespla, fsanchez, mlf}@dlsi.ua.es

Departament de Llenguatges i Sistemes Informàtics
Universitat d’Alacant, E-03071 Alacant, Spain

15th Annual Conference of the EAMT
### Translation Memories

<table>
<thead>
<tr>
<th>English</th>
<th>Catalan</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s_1): European Association for Machine Translation</td>
<td>(t_1): Associació Europea per a la Traducció Automàtica</td>
</tr>
<tr>
<td>(s_2): The EAMT is a member of the IAMT</td>
<td>(t_2): L’EAMT és membre de l’IAMT</td>
</tr>
<tr>
<td>(s_3): current year’s conference is held in Leuven</td>
<td>(t_3): el congrés d’enguany se celebra a Lovaina</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
</tr>
</tbody>
</table>
# Translation Memories

<table>
<thead>
<tr>
<th>English</th>
<th>Catalan</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s_1): European Association for Machine Translation</td>
<td>(t_1): Associació Europea per a la Traducció Automàtica</td>
</tr>
<tr>
<td>(s_2): The EAMT is a member of the IAMT</td>
<td>(t_2): L’EAMT és membre de l’IAMT</td>
</tr>
<tr>
<td>(s_3): current year’s conference is held in Leuven</td>
<td>(t_3): el congrés d’enguany se celebra a Lovaina</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**New sentence** \(s'\): The AMTA is a member of the IAMT
## Translation Memories

<table>
<thead>
<tr>
<th>English</th>
<th>Catalan</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$: European Association for Machine Translation</td>
<td>$t_1$: Associació Europea per a la Traducció Automàtica</td>
</tr>
<tr>
<td>$s_2$: The EAMT is a member of the IAMT</td>
<td>$t_2$: L’EAMT és membre de l’IAMT</td>
</tr>
<tr>
<td>$s_3$: current year’s conference is held in Leuven</td>
<td>$t_3$: el congrés d’enguany se celebra a Lovaina</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**New sentence** $s'$: The AMTA is a member of the IAMT

**Best match** $s_2$: The **EAMT** is a member of the IAMT
## Translation Memories

<table>
<thead>
<tr>
<th>English</th>
<th>Catalan</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$: European Association for Machine Translation</td>
<td>$t_1$: Associació Europea per a la Traducció Automàtica</td>
</tr>
<tr>
<td>$s_2$: The EAMT is a member of the IAMT</td>
<td>$t_2$: L’EAMT és membre de l’IAMT</td>
</tr>
<tr>
<td>$s_3$: current year’s conference is held in Leuven</td>
<td>$t_3$: el congrés d’enguany se celebra a Lovaina</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**New sentence**  $s'$: The AMTA is a member of the IAMT

**Best match**  $s_2$: The EAMT is a member of the IAMT

**Proposal**  $t_2$: L’EAMT és membre de l’IAMT
Fuzzy Matching Scores

Fuzzy matching scores measure the similarity between segments \( s' \) (segment to be translated) and \( s_i \) (matching segment in the Translation memory)

\[
\text{score}(s', s_i) = 1 - \frac{\text{EditDistance}(s', s_i)}{\max(|s'|, |s_i|)}
\]
Fuzzy Matching Scores

Fuzzy matching scores measure the similarity between segments $s'$ (segment to be translated) and $s_i$ (matching segment in the Translation memory)

$$\text{score}(s', s_i) = 1 - \frac{\text{EditDistance}(s', s_i)}{\max(|s'|, |s_i|)}$$

**Example**

$s'$: The Association for Machine Translation in the Americas is the American branch of the IAMT

$s_i$: The European Association for Machine Translation is a member of the IAMT

$$\text{score}(s', s_i) = 1 - \frac{7}{15} \approx 0.53$$
The European Association for Machine Translation is the European branch of the International Association for Machine Translation.

It is a non-profit organisation and organises conferences and workshops on the subject of machine translation.

It was registered in 1991 in Switzerland and is the only organisation of its type in Europe.
Fuzzy Match Scores + Alignment

Edit distance provides information about the matching words between $s'$ and $s_i$:

**Example**

\[
\begin{align*}
\text{t}_i & \quad \text{I’ Associació Europea per a la Traducció Automàtica} \\
\text{s}_i & \quad \underline{\text{the European Association for Machine Translation}} \\
\text{s’} & \quad \underline{\text{the Asia-Pacific Association for Machine Translation}}
\end{align*}
\]
Word alignment may be used to “project” source-side matching information onto $t_i$ to suggest which words to change and which to keep unedited:

**Example**

```
<table>
<thead>
<tr>
<th>$t_i$</th>
<th>l’ Associació Europea per a la Traducció Automàtica</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_i$</td>
<td>the European Association for Machine Translation</td>
</tr>
<tr>
<td>$s'$</td>
<td>the Asia-Pacific Association for Machine Translation</td>
</tr>
</tbody>
</table>
```
Related Work

- **Simard (2003)**: Statistical MT techniques allows exploiting TMs at sub-segment (sub-sentential) level: *translation spotting*

- **Bourdaillet et al. (2009)**: Similar approach for a bilingual concordancer, *TransSearch*

- **Kranias and Samiotou (2004)**: Sub-segment level alignments using a bilingual dictionary to (i) detect words to be changed and (ii) propose translations for them
**Rationale**

- $w_{ij}$ and $v_{ik}$ **aligned** and $v_{ik}$ **matched** $\implies$ **keep** $w_{ij}$
- $w_{ij}$ and $v_{ik}$ **aligned** and $v_{ik}$ **not matched** $\implies$ **change** $w_{ij}$
- $w_{ij}$ **not aligned** $\implies$ ???
Rationale

What to do if there is more than one alignment with contradictory evidence?

\[
\begin{align*}
\text{matched with } s' & \quad \text{unmatched with } s' \\
\end{align*}
\]
We define the likelihood of keeping the word $w_{ij}$ unedited as:

$$f_K(w_{ij}, s', s_i, t_i) = \frac{\sum_{v_{ik} \in \text{aligned}(w_{ij})} \text{matched}(v_{ik})}{|\text{aligned}(w_{ij})|}$$

- $\text{aligned}(w_{ij})$: set of source-side words aligned with $w_{ij}$ in $s_i$
- $\text{matched}(v_{ik})$: 1 if $v_{ik}$ is matched in $s'$ and 0 otherwise
Interpretation of $f_K(w_{ij}, s', s_i, t_i)$

Two ways to interpret $f_K(w_{ij}, s', s_i, t_i)$:

- **Unanimity:**
  - if $f_K(w_{ij}, s', s_i, t_i) = 1$: $w_{ij} \rightarrow$ keep unedited
  - if $f_K(w_{ij}, s', s_i, t_i) = 0$: $w_{ij} \rightarrow$ change
  - otherwise $\rightarrow$ not marked

- **Majority:**
  - if $f_K(w_{ij}, s', s_i, t_i) > \frac{1}{2}$: $w_{ij} \rightarrow$ keep unedited
  - if $f_K(w_{ij}, s', s_i, t_i) < \frac{1}{2}$: $w_{ij} \rightarrow$ change
  - otherwise $\rightarrow$ not marked
Example of Unanimity Criterion

$\text{[change]}$ $\text{[?] [keep] [keep]}$

$t_i$: he missed his brother

$s_i$: él echó de menos a su hermano

$s'_{i}$: ella echó de casa a su hermano
Example of Majority Criterion

$t_i$: he missed his brother

$s_i$: él echó de menos a su hermano

$s’$: ella echó de casa a su hermano
Corpora

- JRC-ACQUIS
  - 5,000 p.o.s.
  - Evaluation Set

- 10,000 p.o.s.
  - Evaluation Translation Memory

Diagram showing the flow of data from JRC-ACQUIS to Evaluation Set and Evaluation Translation Memory.
Evaluation Metrics

Accuracy = \frac{\text{correctly marked words}}{\text{marked words}}

Coverage = \frac{\text{marked words}}{\text{total words}}
We use the GIZA++ (Och and Ney, 2003) free/open-source tool to obtain SL to TL alignment and a TL to SL alignment on the TM. We experiment with three ways to combine the alignments:

- union
- intersection
- grow-diag-final-and
Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood $f_K$ (unanimity or majority)
Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood $f_K$ (unanimity or majority)
- the use of alignment models trained on:
Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood $f_K$ (unanimity or majority)
- the use of alignment models trained on:
  - the corpus to be aligned itself
Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood $f_K$ (unanimity or majority)
- the use of alignment models trained on:
  - the corpus to be aligned itself
  - a separate in-domain corpus
Experimental Settings

We tried our approach comparing:

- the use of three different methods to combine the alignments generated with GIZA++
- the use of both criteria defined to use the likelihood $f_K$ (unanimity or majority)
- the use of alignment models trained on:
  - the corpus to be aligned itself
  - a separate in-domain corpus
  - a separate out-of-domain corpus
Corpora

JRC-ACQUIS

5,000 p.o.s.

EVALUATION SET

10,000 p.o.s.

EVALUATION TRANSLATION MEMORY

STATISTICAL WORD ALIGNMENT TRAINING

Current and future Work
Corpora

- JRC-ACQUIS
  - 5,000 p.o.s.
  - 10,000 p.o.s.
- IN-DOMAIN CORPUS
  - 10,000 p.o.s.
- EVALUATION SET
- EVALUATION TRANSLATION MEMORY
- STATISTICAL WORD ALIGNMENT TRAINING
Corpora

- JRC-ACQUIS
  - 10,000 p.o.s.
  - 5,000 p.o.s. to EVALUATION SET
- EMEA
  - 10,000 p.o.s.
- IN-DOMAIN CORPUS
  - 10,000 p.o.s.
- OUT-OF-DOMAIN CORPUS
  - 10,000 p.o.s.

STATISTICAL WORD ALIGNMENT TRAINING
Results for the Majority/Unanimity Criteria

![Graph showing Accuracy and Coverage](image-url)
Results for the Different Alignment Models

Accuracy (%) vs. Fuzzy-Matching Threshold (%)

Coverage (%) vs. Fuzzy-Matching Threshold (%)
Concluding Remarks

- new method to improve TM-based CAT tools
- predictability and high confidence of translators on fuzzy-match scores is kept
- accuracy over 94% for fuzzy match thresholds between 60% and 90%
- it is possible to reuse statistical alignment models from different corpora with a small loss in accuracy (but a larger loss in coverage)
Current and future Work

Current:
- surveying translators about the usefulness of target-side colouring (visit survey at http://transducens.dlsi.ua.es/people/fsanchez/survey.html)
- using MT to inform aligners and classifiers to colour target words in proposals on the fly (no need to train the aligner on a corpus)

Future:
- integration in the OmegaT free/open-source CAT system
Acknowledgements:
- Partially funded by the Spanish government through project TIN2009-14009-C02-01.
- Good ideas from Yanjun Ma, Andy Way and Harold Somers: thanks!

License: This work may be distributed under the terms of
- the Creative Commons Attribution–Share Alike license:
  http://creativecommons.org/licenses/by-sa/3.0/
- the GNU GPL v. 3.0 License:
  http://www.gnu.org/licenses/gpl.html

Dual license! E-mail me to get the sources: mespla@dlsi.ua.es