On the annotation of TMX translation memories for advanced leveraging in computer-aided translation

Mikel L. Forcada

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

May 30, 2014: Language Resources and Evaluation Conference LREC 2014, Reykjavík, Ísland
1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
A quick review of concepts:

- **Translation memory (TM):** a set of *translation units*
- A *translation unit (TU):* pair of text *segments:*
  - each in a different language
  - mutual translations
- TMs store previous translation jobs in a reusable way.
<table>
<thead>
<tr>
<th>English</th>
<th>Catalan</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 ): The political situation is difficult</td>
<td>( t_1 ): La situació política és difícil</td>
</tr>
<tr>
<td>( s_2 ): The humanitarian situation worsens</td>
<td>( t_2 ): La situació humanitòria em-pitjora</td>
</tr>
<tr>
<td>( s_3 ): Humanitarian efforts have failed</td>
<td>( t_3 ): Els esforços humanitaris han fracassat</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

*Fuzzy matches* of a new sentence \( s' \) help translate it:

**New sentence**  \( s' \): The humanitarian situation is difficult

**Best match**  \( s_2 \): The political situation is difficult

**Proposal**  \( t_2 \): La situació política és difícil

**Edited proposal**  \( t_2 \rightarrow t' \): La situació humanitòria és difícil
Outline

1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
Translation memory exchange (TMX).
- A well established, industry-agreed standard.
- Based on XML
- For the interchange of TMs among computer-aided translation (CAT) applications.

Example of a translation unit in TMX

```xml
<tu segtype="sentence" tuid="2">
  <tuv xml:lang="en">
    <seg>The humanitarian situation worsens.</seg>
  </tuv>
</tu>

<tu segtype="sentence" tuid="2">
  <tuv xml:lang="ca">
    <seg>La situació humanitària empitjora.</seg>
  </tuv>
</tu>
```
Outline

1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
To automate the needed change,\textsuperscript{1} namely,

| New sentence   | \( s' \): The humanitarian situation is difficult |
| Best match     | \( s_2 \): The political situation is difficult |
| Proposal       | \( t_2 \): La situació Política és difícil |
| Edited proposal| \( t_2 \rightarrow t' \): La situació humanitària és difícil |

It would be helpful to know, for instance, that

\[
\text{political situation} \rightarrow \text{situació política} \\
\text{humanitarian situation} \rightarrow \text{situació humanitària}
\]

These sub-segment correspondences are in the TM but they are not annotated.

But they might as well have been!

\textsuperscript{1}This is sometimes called \textit{fuzzy-match repair}.
The term **advanced leveraging**...  
- ...refers to *extensions* beyond current TM usage...  
- ...coming from identifying *sub-segment* repetitions.  

Commercial examples:  
- *Deep Miner* in Atril’s Déjà Vu  
- *Auto-Suggest* in SDL Trados  
- *Advanced Leveraging* in Multicorpora  

TMX does not directly support sub-segment equivalence annotation. Or does it?
Outline

1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
Annotating TMX with sub-segment information

After considering some alternatives (see paper):

- **Proposal:** repurposing existing support in TMX for *overlapping format paired tags* (yuck!)

### Overlapping paired format tags in English

<B>Bold, <I>Bold + Italic</I>, Italic</B>.

### Corresponding (also overlapping) paired format tags in Spanish

<B>Negrita, <I>Negrita + Cursiva</I>, Cursiva</B>.

In TMX, one can

- Use an index *i* to pair each *begin paired tag* (**bpt**) with the corresponding *end paired tag* (**ept**) in the same segment.
- Use an index *x* to align each tag in one language with the corresponding tag in the other language.
<tu segtype="sentence" tuid="877">
    <tuv xml:lang="en">
        <seg>
            <bpt i="1" x="1">Bold, </bpt>
            <bpt i="2" x="2">Bold + Italic</bpt>.
        </seg>
    </tuv>
    <tuv xml:lang="es">
        <seg>
            I have written
            Negrita, Negrita + Cursiva.
        </seg>
    </tuv>
</tu>
Annotating TMX with sub-segment information

The solution:² *null (empty) format tags.* In TMX:

- Each `<ept>`–`<bpt>` pair may clearly span any arbitrary subsegment in *seg*
- Elements `<ept>` and `<bpt>` *can be empty!*
- An attribute type may be used to specify “the kind of data [the] element represents”

Therefore

- We can use aligned `<ept>`–`<bpt>` pairs *containing no format* to represent subsegment correspondences
- We can *twist* the accepted use of the type attribute to encode the *source of information* used to annotate that correspondence.

²thanks Felipe Sánchez-Martínez!
TMX translation unit with one subsegment annotated

```xml
<tu segtype="sentence" tuid="13123123">
  <tuv xml:lang="de">
    <seg>Ich habe einen Artikel geschrieben.</seg>
  </tuv>
  <tuv xml:lang="en">
    <seg>I have written an article.</seg>
  </tuv>
</tu>
```
Annotating TMX with sub-segment information

TMX translation unit with two overlapping subsegments annotated

```xml
<tu segtype="sentence" tuid="13123123">
  <tuv xml:lang="de">
    <seg>Ich
      <bpt i="1" x="1" type="google-translate-de-en"/>gehe
      <bpt i="2" x="2" type="google-translate-de-en"/>ins
      <ept i="1"/> Haus<ept i="2"/>.
    </seg>
  </tuv>
  <tuv xml:lang="en">
    <seg>I
      <bpt i="1" x="1" type="google-translate-de-en"/>go
      <bpt i="2" x="2" type="google-translate-de-en"/>into the
      <ept i="1"/> house<ept i="2"/>.
    </seg>
  </tuv>
</tu>
```
Pros and cons of \texttt{<ept>} and \texttt{<bpt>} repurposing.

Pros:

- This method allows for a very general annotation of all kinds of subsegment correspondences.
- A related localization standard, XLIFF, also uses \texttt{<ept>} and \texttt{<bpt>} with similar syntax and semantics.
  - It remains to be seen if it would be possible to \textit{twist} XLIFF too!

Cons:

- Extending the semantics of \texttt{<bpt>} and \texttt{<ept>} could give trouble with CAT systems that explicitly consider them (instead of just stripping them)
- Does not explicitly encode sub-segment correspondences as separate translation units \texttt{<tu>} (always bound to a subsegment, may be repeated somewhere else).

In statistical machine translation parlance, one would say that “the phrase table is embedded in the \textit{bilingual training corpus}”. 
Outline

1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
Sources of subsegment equivalence

Subsegment equivalences may come from...

- ...smaller translation units in the same TM or another TM.
- ...an external source of bilingual equivalence such as a machine translation system...
  - note that in this case, MT output is “validated” by the existing translation in the translation memory
- ...or a term base.
- ...a statistical word alignment of the current translation memory.
  - subsegment pairs can be those compatible with those word alignments.
Outline

1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
Concluding remarks

- I have presented a proposal\(^3\) to enrich TMX-encoded translation memories with information about subsegment equivalence
  - Ready for *advanced leveraging*
- It repurposes existing resources for formatting in the TMX standard
- Subsegment annotation may be *generated in advance* using
  - Machine translation
  - [Statistical] word alignment followed by *phrase-pair* extraction
  - Smaller TUs from the same or other TMs
  - Term bases, glossaries, etc.
- and *stored* together with the TMX file.

\(^3\)The *paper* discusses other alternatives
Thank you!

Mikel L. Forcada

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

Support from the Spanish Ministry of Economy and Competitiveness through grant TIN2012-32615 is gratefully acknowledged.

I also thank Felipe Sánchez-Martínez and Juan Antonio Pérez-Ortiz for interesting suggestions.

Finally, I thank Google Summer of Code student Pankaj Kumar Sharma for experimental implementations using Apertium to annotate subsegments in a TMX memory.
This slide has been intentionally left empty
1. Computer-aided translation using translation memories
2. The TMX standard
3. The need for sub-segment annotation: advanced leveraging
4. A proposal for sub-segment correspondence annotation in TMX
5. Sources of sub-segment equivalence
6. Concluding remarks
7. [Spare slides: other alternatives considered]
A possibility uses `<prop>` (“used to define properties of the parent element”), storing sub-segments as separate `<tu>` (“stand-off”)?:

The annotating subsegment TU specifies how it annotates a TU

```xml
<tu segtype="phrase" tuid="984120312">
  <prop type="annotated-tuid">13123123</prop>
  <prop type="source">google-translate-de-en</prop>
  <tuv xml:lang="de">
    <prop type="start-pos">10</prop>
    <prop type="end-pos">22</prop>
    <seg>einen Artikel</seg>
  </tuv>
  <tuv xml:lang="en">
    <prop type="start-pos">16</prop>
    <prop type="end-pos">25</prop>
    <seg>an article</seg>
  </tuv>
</tu>
```
Discarded alternative: using `<prop>/2

- Treats sub-segment correspondences as TUs (natural).
- Cumbersome `<prop>` overloading for common sub-segment pairs
- Use of character offsets may be fragile
- Matching `<prop>` lists would be needed in annotated TUs:

The annotated TU names the annotating sub-segment TUs

```xml
<tu segtype="sentence" tuid="13123123">
  <prop type="annotated-by-tuid">984120312</prop>
  <tuv xml:lang="de">
    <seg>Ich habe einen Artikel geschrieben.</seg>
  </tuv>
  <tuv xml:lang="en">
    <seg>I have written an article</seg>
  </tuv>
</tu>
```
Discarded alternative: using \texttt{<hi>/1}

A possibility would use \texttt{<hi>} ("used to delimit a portion of the segment for any user-defined purpose"): 

\begin{verbatim}
<tu segtype="sentence" tuid="13123123">
  <tuv xml:lang="de">
    <seg>Ich habe
      <hi x="1" type="google-translate-de-en">einen
      Artikel</hi> geschrieben.</seg>
  </tuv>
  <tuv xml:lang="en">
    <seg>I have written
      <hi x="1" type="google-translate-de-en">an
      article</hi></seg>
  </tuv>
</tu>
\end{verbatim}
Discarded alternative: using `<hi>`/2

- Allows for a rather rich annotation of sub-segment correspondence without having to stretch too far the intended semantics of the `<hi>` element.
- Element `<hi>` may be indefinitely nested, but no overlap is possible.
- It may however be OK if a clear phrase structure is defined (for instance using a synchronous context-free grammar):

  
  \[
  \begin{align*}
  [1] & \text{Ich} \quad [2] \text{habe} \quad [3] [4] \text{einen Artikel} \quad \text{geschrieben} \quad ] \quad ] \\
  [1] & \text{I} \quad [2] \text{have} \quad [3] \text{written} \quad [4] \text{an article} \quad ] \quad ]
  \end{align*}
  \]