Target-language edit hints: a basic description of the method[*]

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, Mikel L. Forcada
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant
{mespla,fsanchez,mlf}@dlsi.ua.es

October 24, 2011

1 Introduction

This document is aimed at giving a basic idea about how the *simple* method in the *target-language edit hints* extension for OmegaT works. This method uses machine translation to detect which target-language words should be kept and which of them should be changed. It works on the fly when a matching translation unit (s, t) is found for a given source-language segment s' to be translated. We will describe this method by means of an example which will be developed step by step along this document. For our example, we suppose a segment

s'=*Costarà dies solucionar el cas*

to be translated from Catalan to English by means of a translation memory which proposes a matching translation unit (s, t) with

s=*Costarà temps solucionar el problema*
t=*It will take time to solve the problem.*

A possible translation t' for s' using t as a base may be t'=*It will take days to solve the case.*

2 First step: Sub-segment alignment

Our method uses machine translation in order to align the sub-segments of both segments s and t in a matching translation unit (s, t). From these alignments the information about which words to change and which of them to keep untouched is derived. It is worth noting that our method is not actually used to translate any new material for the user and that no translated text is ever shown to the suers of OmegaT.

To obtain the sub-segment alignments, both the source-language segment s and the target-language segment t in each translation unit are segmented in sub-segments of length between 1 and 3 words.[1] Let σ be a sub-segment from s and τ a sub-segment from t. We consider that σ and τ are aligned if any of the available source of bilingual information confirm that σ is a translation of τ, or vice-versa.

In the example we proposed, we may obtain the following set of sub-segment alignments by using Apertium as the machine translator:

temps	↔	*time*
problema	↔	*problem*
solucionar el	↔	*solve the*
el problema	↔	*the problem*
solucionar el problema	↔	*solve the problem*

Figure 1 shows a graphical representation of these alignments.

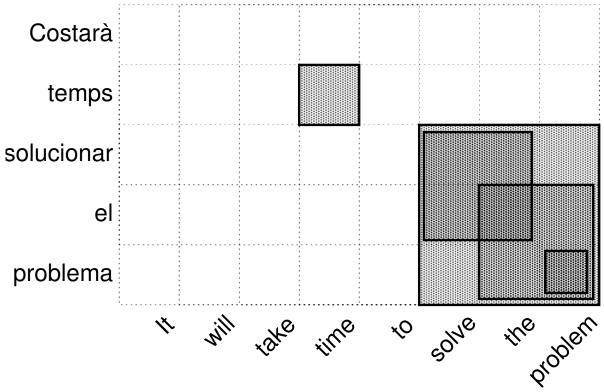


Figure 1: Sub-segment alignments.

[1] Longer sub-segments may be used, but we chose these lengths because no significant improvement was obtained by using longer sub-segments.

It is worth noting that, at this point, multiple machine translation systems may be used in order to obtain as many sub-segment alignments as possible. It could be also possible to use other bilingual information sources, such as glossaries or other translation memories, but, for the moment, the extension for OmegaT only supports machine translation.

3 Second step: Colouring words with sub-segment alignments information

We use the information provided by the sub-segment alignments to decide which target words in t should be kept untouched and which words should be edited. To do so, we compute the alignment strength S_{jk} between the j-th word in s and the k-th word in t as

$$S_{jk}(s,t,M) = \sum_{(\sigma,\tau)\in M} \frac{\mathrm{cover}(j,k,\sigma,\tau)}{|\sigma|\cdot|\tau|}$$

where M is the set of sub-segment alignments detected for translation unit (s,t), $|x|$ is the length of segment x measured in words, and $\mathrm{cover}(j,k,\sigma,\tau)$ equals 1 if σ covers the j-th word in s and τ the k-th word in t, and 0 otherwise. This way of computing the alignment strengths is based on the idea that sub-segment alignments apply *alignment pressures* on the words; so the bigger surface covered by the sub-segment alignment, the lower the word-alignment strength obtained. Following our example, the alignment strengths for the words covered by sub-segment alignments are presented in Figure 2. The words *temps* and *time* are only covered by a sub-segment alignment (*temps*,*time*), so the surface is 1 and the alignment strength is $S_{1,4} = 1$. However, words *problema* and *problem* are covered by three sub-segment alignments: (*problema*, *problem*), (*el problema*, *the problem*), and (*solucionar el problema*, *solve the problem*). So the alignment strength is $S_{3,6} = \frac{1}{1} + \frac{1}{4} + \frac{1}{9} = 1.36$.

The alignment strengths can then be used to decide which words should be kept unedited and which of them should be changed. We define the likelihood $L_k(s',s,t,M)$ that the k-th word in the segment t should be kept unedited as:

$$L_k(s',s,t,M) = \frac{\sum_{j=0}^{|s|} S_{jk}(s,t,M)\cdot \mathrm{matched}(j,s,s')}{\sum_{j=0}^{|s|} S_{jk}(s,t,M)}$$

where $\mathrm{matched}(j,s,s')$ equals 1 if the j-th word in s is matched in s' and 0 otherwise. Our method recommends to keep untouched words with

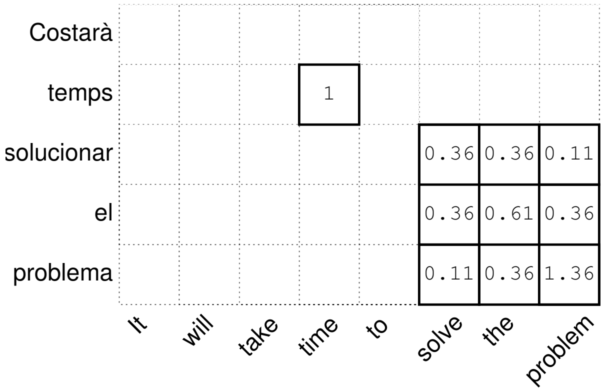
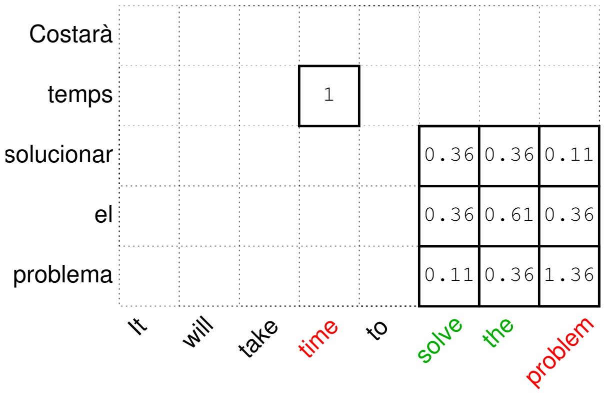


Figure 2: Alignment strengths.

$L_k(s', s, t, M) \geq 0.5$ and to change words with $L_k(s', s, t, M) < 0.5$.[2] No recommendation is made for target-language words which do not appear in any sub-segment alignment, since no information is available for making any recommendations about them. Following with the example proposed, Figure 3 shows the recommendations made by our method for the words in target-language segment t. Source-language words in green represent those words which are matched between s' and s and words in red represent those source-language words which are unmatched. As can be seen, target-language words *time* and *problem* are marked for editing, since their likelihood is $L_3(s', s, t, M) = \frac{0}{1} = 0$ and $L_7(s', s, t, M) = \frac{0.47}{1.83} = 0.26$, respectively. Conversely, target-language words *solve*, and *the* are recommended to be kept, since they have values higher than 0.5 for the keeping likelihood: $L_5(s', s, t, M) = \frac{0.72}{0.83} = 0.87$ and $L_6(s', s, t, M) = \frac{0.97}{1.33} = 0.73$, respectively. Finally, no recommendation is done for target-language words *It*, *will*, and *take*, since they do not appear in any sub-segment alignment. It is worth noting that the recommendations made by this extensions are suitable for the translation t'=*It will take days to solve the case* proposed for the segments$'$.

[2] This could be adjusted for performance in the future



	It	will	take	time	to	solve	the	problem
Costarà								
temps				1				
solucionar						0.36	0.36	0.11
el						0.36	0.61	0.36
problema						0.11	0.36	1.36

Figure 3: Obtaining target-language edit hints.