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

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

Abstract

This paper explores a new method to improve computer-aided translation (CAT) systems based on translation memory (TM) by using pre-computed word alignments between the source and target segments in the translation units (TUs) of the user's TM. When a new segment is to be translated by the CAT user, our approach uses the word alignments in the matching TUs to mark the words that should be changed or kept unedited to transform the proposed translation into an adequate translation. In this paper, we evaluate different sets of alignments obtained by using GIZA++. Experiments conducted in the translation of Spanish texts into English show that this approach is able to predict which target words have to be changed or kept unedited with an accuracy above 94% for fuzzy-match scores greater or equal to 60%. In an appendix we evaluate our approach when new TUs (not seen during the computation of the word-alignment models) are used.

1 Introduction

Computer-aided translation (CAT) systems based on translation memory (TM) (Bowker, 2002; Somers, 2003) and, optionally, additional tools such as terminology databases (Bowker, 2003), are the translation technology of choice for most professional translators, especially when translation tasks are very repetitive and effective recycling of previous translations is feasible.

When using a TM-based CAT system to translate a source segment s', the system provides the set of translation units (TUs) $\{(s_i, t_i)\}_{i=1}^{N}$ whose fuzzy-match score is above a given threshold Θ, and marks which words in each source-language (SL) segment s_i differ from those in s'. It is however up to the translator to identify which target words in the corresponding target-language (TL) segments t_i should be changed to convert t_i into t', an adequate translation of s'.

The method we propose and evaluate in this paper is aimed at recommending the CAT user which words of t_i should be changed by the translator or kept unedited to transform t_i into t'. To do so, we pre-process the user's TM to compute the word alignments between the source and target segments in each TU. Then, when a new segment s' is to be translated, the TUs with a fuzzy-match score above the threshold Θ are obtained and the alignment between the words in s_i and t_i are used to mark which words in t_i should be changed or kept unedited.

Related work. In the literature one can find different approaches that use word or phrase alignments to improve existing TM-based CAT systems; although, to our knowledge, none of them use word alignments for the purpose we study in this paper. Simard (2003) focuses on the creation of TM-based CAT systems able to work at the sub-segment level by proposing as translation sub-segments extracted from longer segments in the matching TUs. To do this, he implements the *translation spotting* (Véronis and Langlais, 2000) technique by using statistical word-alignment methods (Och and Ney, 2003); translation spotting consists of identifying, for a pair of parallel sentences, the words or phrases in a TL segment that correspond to the words in a SL segment. The work by Bourdaillet et al. (2009) follows a similar approach, although it does not focus on traditional TM-based CAT systems, but

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 81–88
Leuven, Belgium, May 2011

on the use of a bilingual concordancer to assist professional translators.

More similar to our approach is the one by Kranias and Samiotou (2004) which is implemented on the *ESTeam* CAT system. Kranias and Samiotou (2004) align the source and target segments in each TU at different sub-segment levels by using a bilingual dictionary (Meyers et al., 1998), and then use these alignments to (i) identify the sub-segments in a translation proposal t_i that need to be changed, and (ii) propose a machine translation for them.

In this paper we propose a different way of using word alignments in a TM-based CAT system to alleviate the task of professional translators. The main difference between our approach and those previously described is that in our approach word alignments are used only to recommend the words to be changed or kept unedited, without proposing a translation for them, so that the user can focus on choosing a translation where words have to be changed. It is worth noting that as we do not change the translation proposals in any way, our approach does not affect the predictability of TM proposals and the way in which fuzzy-match scores (Sikes, 2007) are interpreted by the CAT user. In addition, our system is independent of any external resources, such as MT systems or dictionaries, as opposed to the work by Kranias and Samiotou (2004).

The rest of the paper is organized as follows. Section 2 presents the way in which word alignments are used by our approach and the different word alignment methods we have tried. Section 3 then describes the experimental framework, whereas Section 4 discusses the results obtained. Section 5 includes some concluding remarks and plans for future work. In Appendix A we evaluate our approach when it is applied to new TUs not seen during the computation of the word-alignment models used.

2 Methodology

Let w_{ij} be the word in the j-th position of segment t_i which is aligned with word v_{ik} in the k-th position of its counterpart segment s_i. If v_{ik} is part of the match between s_i and s' (the new segment to be translated), then this indicates that w_{ij} might be part of the translation of that word and, therefore, it should be kept unedited. Conversely, if $v_{ik'}$ is not part of the match between s_i and s', this indicates that $w_{ij'}$ might not be the translation of any of the words in s' and it should be changed (see Figure 1). Note that w_{ij} may not be aligned with any word

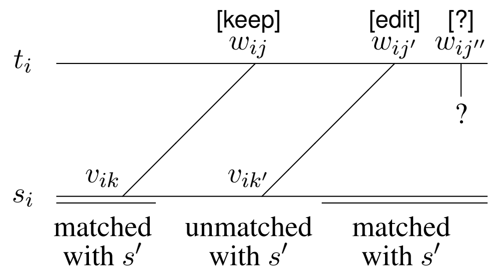

Figure 1: Target word w_{ij} may have to be kept unedited because it is aligned with source word v_{ik} which is in the part of s_i that matches s'. Target word $w_{ij'}$ may have to be changed because it is aligned with source word $v_{ik'}$ which is in the part of s_i that does not match s'. As target word $w_{ij''}$ is not aligned to any source word in s_i, nothing can be said about it.

in s_i, and that in these cases nothing can be said about it. This information may be shown using colour codes, for example, red for the words to be changed, green for the words to be kept unedited and yellow for those unaligned words for which nothing can be said.

To determine if word w_{ij} in the target proposal t_i should be changed or kept unedited, we compute the fraction of words v_{ik} aligned to w_{ij} which are common to both s_i and s':

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

where $\text{aligned}(w_{ij})$ is the set of source words in s_i which are aligned with target word w_{ij} in t_i, and $\text{matched}(v_{ik})$ equals 1 if word v_{ik} is part of the match between s_i and s', the segment to be translated, and 0 otherwise. Function $\text{matched}(x)$ is based on the optimal edit path, obtained as a result of the word-based Levenshtein distance (Levenshtein, 1966) between the segment to be translated and the SL segment of the matching TU. The fraction $f_K(w_{ij}, s', s_i, t_i)$ may be interpreted as the likelihood that word w_{ij} has to be kept unedited. If $|\text{aligned}(w_{ij})|$ happens to be zero, $f_K(w_{ij}, s', s_i, t_i)$ is arbitrarily set to $\frac{1}{2}$, meaning "do not know".

We have chosen the likelihood that word w_{ij} will be kept unedited to depend on how many SL words aligned with it are matched with the SL segment to be translated. It may happen that w_{ij} is aligned with one or more words in s_i that are matched with words in s', and, at the same time, it is aligned with one or more unmatched words in s_i. In the experiments we have tried two ways of dealing with this, one that requires all SL words in s_i to be matched, and another one that only requires the majority of words aligned with w_{ij} to be matched. These strategies have been chosen because of their simplicity, although it could also be possible to use, for example, a maximum entropy classifier (Berger et al., 1996), in order to determine which words should be changed or kept unedited. In that case, f_K would be one of the features used by the maximum entropy classifier.

To illustrate these ideas, Figure 2 shows an example of a word-aligned pair of segments (s_i and t_i) and a segment s' to be translated. As can be seen, the word *he* in t_i is aligned with the word *él* in s_i, which does not match with any word in s'. Therefore, *he* should be marked to be changed. Conversely, the words *his* and *brother* are aligned with *su* and *hermano*, respectively, which are matched in s' and, therefore should be kept unedited. Finally, the word *missed* is aligned with three words in s_i: *echó* and *de*, which are matched in s', and *menos*, which is not matched. In this case, if the criterion of unanimity is applied, the word would be marked neither as "keep" nor as "change". Otherwise, if the criterion of majority is applied, the word would be marked to be changed.

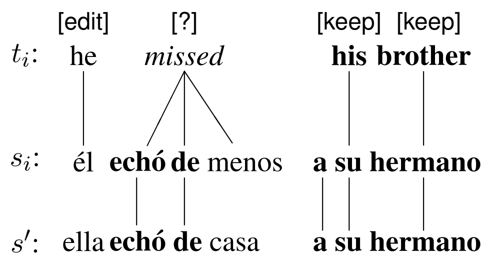

Figure 2: Example of alignment and matching.

For the experiments in this paper we have used word alignments obtained by means of the free/open-source GIZA++[1] tool (Och and Ney, 2003) which implements standard word-based statistical machine translation models (Brown et al., 1993) as well as a hidden-Markov-model-based alignment model (Vogel et al., 1996). GIZA++ produces alignments in which a source word can be aligned with many target words, whereas a target word is aligned with, at most, one source word. Following common practice in statistical machine translation (Koehn, 2010, Ch. 4) we have obtained a set of symmetric word alignments by running GIZA++ in both translation directions, and then symmetrizing both sets of alignments. In the experiments we have tried the following symmetrization methods:

- the union of both sets of alignments,
- the intersection of the two alignment sets, and
- the use of the *grow-diag-final-and* heuristic (Koehn et al., 2003) as implemented in Moses (Koehn et al., 2007).

3 Experimental settings

We have tested our approach in the translation of Spanish texts into English by using two TMs: $\mathrm{TM}_{\mathrm{trans}}$ and $\mathrm{TM}_{\mathrm{test}}$. Evaluation was carried out by simulating the translation of the SL segments in $\mathrm{TM}_{\mathrm{trans}}$ by using the TUs in $\mathrm{TM}_{\mathrm{test}}$. We firstly obtained the word alignments between the parallel segments of $\mathrm{TM}_{\mathrm{test}}$ by training and running GIZA++ on the TM itself. Then, for each source segment in $\mathrm{TM}_{\mathrm{trans}}$, we obtained the TUs in $\mathrm{TM}_{\mathrm{test}}$ having a fuzzy-match score above threshold Θ, and tagged the words in their target segments as "keep" or "change".

3.1 Fuzzy-match score function

As in most TM-based CAT systems, we have chosen a fuzzy-match score function based on the Levenshtein distance (Levenshtein, 1966):

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

where $|x|$ stands for the length (in words) of string x and $\mathrm{D}(x, y)$ refers to the word-based Levenshtein distance (edit distance) between x and y.

3.2 Corpora

The TMs we have used were extracted from the JRC-Acquis corpus version 3 (Steinberger et al., 2006),[2] which contains the total body of European Union (EU) law. Before extracting the TMs used, this corpus was tokenized and lowercased, and then segment pairs in which either of the segments was empty or had more than 9 times words than its counterpart were removed. Finally, segments longer than 40 words (and their corresponding counterparts) were removed because of the inability of GIZA++ to align longer segments.

[1] http://code.google.com/p/giza-pp/

[2] http://wt.jrc.it/lt/Acquis/

$\Theta(\%)$	TUs	N_{words}
50	9.5	484,523
60	6.0	303,193
70	4.5	220,304
80	3.5	166,762
90	0.9	42,708

Table 1: Average number of matching TUs per segment and number of words to tag for different fuzzy-match score thresholds (Θ).

Finally, the segment pairs in $\mathrm{TM}_{\text{trans}}$ and $\mathrm{TM}_{\text{test}}$ were randomly chosen without repetition from the resulting corpus. $\mathrm{TM}_{\text{test}}$ consists of 10,000 parallel segments, whereas $\mathrm{TM}_{\text{trans}}$ consists of 5,000 segment pairs. It is worth noting that these TMs may contain incorrect TUs as a result of wrong segment alignments and this can negatively affect the results obtained.

With respect to the number of TUs found in $\mathrm{TM}_{\text{test}}$ when simulating the translation of the SL segments in $\mathrm{TM}_{\text{trans}}$, Table 1 reports for the different fuzzy-match score thresholds we have used: the averaged number of TUs per segment to be translated and the total number of words to classify as "keep" or "change". These data provide an idea of how repetitive the corpora we have used to carry out the experiments are.

3.3 Evaluation

We evaluate our approach for different fuzzy-match score thresholds Θ by computing the accuracy, i.e. the percentage of times the recommendation of our system is correct, and the coverage, i.e. the percentage of words for which our system is able to say something. For that purpose we calculate the optimal edit path between the target segments in $\mathrm{TM}_{\text{trans}}$ and the translation proposals in $\mathrm{TM}_{\text{test}}$ to determine the actual word-editing needs in each translation proposal.

For each SL segment s' in $\mathrm{TM}_{\text{trans}}$ we compute the set of matching TUs $\{(s_i, t_i)\}_{i=1}^{N}$ in $\mathrm{TM}_{\text{test}}$ whose fuzzy-match score is above threshold Θ. We then calculate the fraction $f_K(w_{ij}, s', s_i, t_i)$ representing the likelihood that word w_{ij} in t_i will be kept unedited and use it to mark w_{ij} as having to be changed or kept unedited by using the two different criteria (unanimity or majority) mentioned above:

unanimity: if $f_K(\cdot) = 1$ the word is tagged as "keep", whereas if $f_K(\cdot) = 0$ it is tagged as "change"; in the rest of cases no recommendation is made for that word.

majority: if $f_K(\cdot) > 0.5$ the word is tagged as "keep", whereas it is tagged as "change" if $f_K(\cdot) < 0.5$; in the unlikely case of having $f_K(\cdot) = 0.5$ no recommendation is made about that word.

The first criterion requires all the source words aligned with word w_{ij} to be matched (conversely, unmatched) with a word in the new segment to be translated, while the second criterion only requires the majority or source words aligned with w_{ij} to be matched (conversely, unmatched).

4 Results and discussion

We evaluated our approach with the different sets of word alignments obtained through the symmetrization methods described in Section 2 for values of the fuzzy-match score threshold Θ between 50% and 90%.

Tables 2 and 3 reports the accuracy and the coverage obtained with each set of alignments together with their confidence intervals for a statistical significance level $p = 0.99$ (DeGroot and Schervish, 2002, Sec. 7.5) when the majority criterion and the unanimity criterion, respectively, are used to mark the words as "keep" or "change".

As can be seen, with both criteria the best accuracy is achieved with the set of alignments obtained through the intersection method, although the use of this set of alignments shows a smaller coverage as compared to the other two sets of alignments. The use of either the union or the grow-diag-final-and sets of alignments seems to have a small impact on the accuracy although the coverage obtained for the union is slightly better. Note that with the alignments obtained by means of the intersection method, both criteria are equivalent because each word is aligned at most with one word in the other language.

The use of the unanimity criterion causes the accuracy to grow slightly, as compared to the majority criterion, while the coverage gets slightly worse as expected. It is worth noting that for fuzzy-match score thresholds above 50% differences in accuracy between both criteria are insignificant, whereas the differences in coverage are small, but significant for values of 60% and 70% of Θ.

Finally, it is important to remark that for values of Θ greater or equal to 60%, which are the values that professional translators tend to use (Bowker,

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	92.35 ± .10	97.33 ± .06	93.80 ± .10	90.78 ± .11	92.34 ± .10	96.73 ± .07
60	94.62 ± .11	98.06 ± .07	95.80 ± .10	92.44 ± .12	94.72 ± .11	97.70 ± .07
70	97.19 ± .10	98.69 ± .06	98.04 ± .08	94.03 ± .13	97.31 ± .09	98.37 ± .07
80	98.31 ± .08	99.05 ± .06	98.82 ± .07	95.50 ± .13	98.44 ± .08	98.78 ± .07
90	97.97 ± .18	99.24 ± .11	98.75 ± .14	95.41 ± .26	98.25 ± .16	98.75 ± .14

Table 2: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the majority criterion for the three different sets of word alignments: intersection, union and grow-diag-final-and (GDFA).

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	92.53 ± .10	96.87 ± .06	93.80 ± .10	90.78 ± .11	92.43 ± .10	96.50 ± .07
60	94.73 ± .11	97.78 ± .07	95.80 ± .10	92.44 ± .12	94.76 ± .11	97.57 ± .07
70	97.26 ± .10	98.50 ± .07	98.04 ± .08	94.03 ± .13	97.35 ± .09	98.30 ± .07
80	98.35 ± .08	98.96 ± .06	98.82 ± .07	95.50 ± .13	98.45 ± .08	98.75 ± .07
90	98.02 ± .18	99.17 ± .11	98.75 ± .14	95.41 ± .26	98.26 ± .16	98.73 ± .14

Table 3: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the unanimity criterion for the three different sets of word alignments: intersection, union and grow-diag-final-and (GDFA).

2002, p. 100), with the three sets of alignments and with both criteria accuracy is always above 94%.

5 Concluding remarks

In this paper we have presented and evaluated a new approach to guide TM-based CAT users by recommending the words in a translation proposal that should be changed or kept unedited. The method we propose requires the TM to be pre-processed in advance in order to get the alignment between the words in the source and target segments of the TUs. In any case, this pre-processing needs to be done only once, although to consider new TUs created by the user it may be worth to re-run the alignment procedure (see Appendix A). The experiments conducted in the translation of Spanish texts into English show an accuracy above 94% for fuzzy-match score thresholds greater or equal to 60% and above 97% for fuzzy-match score thresholds above 60%.

Our approach is intended to guide the TM-based CAT user in a seamless way, without distorting the known advantages of the TM-based CAT systems, namely, high predictability of the translation proposals and easy interpretation of fuzzy-match scores. We plan to field-test this approach with professional translators in order to measure the possible productivity improvements. To do this we will integrate this method in OmegaT,[3] a free/open-source TM-based CAT system.

A Adding new TUs to the TM

In our experiments, we obtained the word alignment models from $\mathrm{TM}_{\mathrm{test}}$ and used them to align the words in the TUs of the same TM. In this way, we used the most information available to obtain the best word alignments possible. However, TMs are not always static and new TUs can be added to them during a translation job. In this case, the previously computed alignment models could be less effective to align the segments in the new TUs.

In this appendix, we evaluate the re-usability of previously computed alignment models on new TUs for our approach. To do so, we used an in-domain TM ($\mathrm{TM}_{\mathrm{in}}$) and an out-of-domain TM ($\mathrm{TM}_{\mathrm{out}}$) to obtain the alignment models and used them to align the segments in the TUs of $\mathrm{TM}_{\mathrm{test}}$. We then repeated the same experiments described in Section 3.3 in order to compare the results obtained.

$\mathrm{TM}_{\mathrm{in}}$ was built with 10,000 pairs of segments extracted from the JCR-Acquis corpus. These pairs of segments were chosen so as to avoid any common TU between $\mathrm{TM}_{\mathrm{in}}$ and $\mathrm{TM}_{\mathrm{test}}$, or between

[3] http://www.omegat.org

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	91.95 ± .10	94.03 ± .09	93.44 ± .10	87.19 ± .12	92.10 ± .10	93.42 ± .09
60	94.34 ± .11	94.06 ± .11	95.53 ± .10	88.26 ± .15	94.51 ± .11	93.74 ± .11
70	97.05 ± .10	93.99 ± .13	97.86 ± .08	89.23 ± .17	97.21 ± .09	93.71 ± .13
80	98.22 ± .09	93.64 ± .15	98.74 ± .07	90.05 ± .19	98.35 ± .08	93.42 ± .16
90	97.88 ± .19	93.61 ± .31	98.69 ± .15	89.81 ± .38	98.10 ± .18	93.28 ± .31

Table 4: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the majority criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM_{in}}$.

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	92.07 ± .10	93.70 ± .09	93.44 ± .10	87.19 ± .12	92.16 ± .10	93.25 ± .09
60	94.39 ± .11	93.87 ± .11	95.53 ± .10	88.26 ± .15	94.53 ± .11	93.66 ± .11
70	97.07 ± .10	93.87 ± .13	97.86 ± .08	89.23 ± .17	97.22 ± .09	93.66 ± .13
80	98.22 ± .09	93.60 ± .15	98.74 ± .07	90.05 ± .19	98.35 ± .08	93.42 ± .16
90	97.88 ± .19	93.60 ± .31	98.69 ± .15	89.81 ± .38	98.10 ± .18	93.27 ± .31

Table 5: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the unanimity criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM_{in}}$.

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	90.57 ± .12	88.03 ± .12	93.83 ± .10	77.13 ± .16	90.37 ± .12	88.27 ± .12
60	93.66 ± .12	88.50 ± .15	96.04 ± .10	79.88 ± .19	93.64 ± .12	88.45 ± .15
70	96.77 ± .10	88.77 ± .17	98.34 ± .08	82.48 ± .21	96.87 ± .10	88.53 ± .18
80	98.10 ± .09	88.29 ± .20	98.96 ± .06	84.39 ± .23	98.23 ± .09	88.05 ± .21
90	97.86 ± .19	90.71 ± .36	98.87 ± .14	84.98 ± .45	98.15 ± .18	90.24 ± .37

Table 6: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the majority criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM_{out}}$.

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	91.15 ± .11	87.22 ± .12	93.83 ± .10	77.13 ± .16	90.87 ± .11	87.74 ± .12
60	93.94 ± .12	88.10 ± .15	96.04 ± .10	79.88 ± .19	93.88 ± .12	88.20 ± .15
70	96.94 ± .10	88.54 ± .18	98.34 ± .08	82.48 ± .21	97.02 ± .10	88.40 ± .18
80	98.16 ± .09	88.22 ± .20	98.96 ± .07	84.39 ± .23	98.29 ± .09	87.99 ± .21
90	97.89 ± .19	90.68 ± .36	98.87 ± .14	84.98 ± .45	98.17 ± .18	90.22 ± .37

Table 7: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the unanimity criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM_{out}}$.

$\mathrm{TM}_{\mathrm{in}}$ and $\mathrm{TM}_{\mathrm{trans}}$. $\mathrm{TM}_{\mathrm{out}}$ was built with 10,000 pairs of segments extracted from the EMEA corpus version 0.3 (Tiedemann, 2009),[4] which is a compilation of documents from the European Medicines Agency, and, therefore, it clearly belongs to a different domain. Before extracting the TUs, the EMEA corpus was pre-processed in the same way that the JRC-Acquis was (see Section 3.2).

Tables 4 and 5 show the results of the experiments when using the alignment models learned from $\mathrm{TM}_{\mathrm{in}}$ for the majority criterion and for the unanimity criterion, respectively. Analogously, tables 6 and 7 show the analogous results when the alignment models learned from $\mathrm{TM}_{\mathrm{out}}$ are used.

As can be seen, the accuracy obtained by our approach when re-using alignment models from an in-domain corpus is very similar to that obtained when these alignments are learned from the TM whose TUs are aligned. Even when the alignment models are learned from an out-of-domain corpus, the loss of accuracy is, in the worst case, lower than 2%. The main problem is the loss of coverage, which is about 6% for the in-domain training and higher than a 10% for the out-of-domain training.

On the one hand, these results show that our approach is able to re-use alignment models computed for a TM on subsequently added TUs keeping a reasonable accuracy in the recommendations. On the other hand, it is obvious that our method becomes less informative for these new TUs as their domain differs from the domain from which the alignment models have been learned.

Acknowledgements: Work supported by Spanish government through project TIN2009-14009-C02-01. The authors thank Yanjun Ma, Andy Way and Harold Somers for suggestions.

References

Berger, A.L., V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bourdaillet, J., S. Huet, F. Gotti, G. Lapalme, and P. Langlais. 2009. Enhancing the bilingual concordancer TransSearch with word-level alignment. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence*, volume 5549 of *Lecture Notes in Artificial Intelligence*, pages 27–38. Springer.

Bowker, L., 2002. *Computer-aided translation technology: a practical introduction*, chapter Translation-Memory Systems, pages 92–127. University of Ottawa Press.

Bowker, L. 2003. Terminology tools for translators. In Somers, H., editor, *Computers and Translation: A Translator's Guide*, pages 49–65. John Benjamins.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

DeGroot, M. H. and M. J. Schervish. 2002. *Probability and Statistics*. Addison-Wesley, third edition.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge University Press.

Kranias, L. and A. Samiotou. 2004. Automatic translation memory fuzzy match post-editing: A step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 52–57, Lisbon, Portugal.

Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.*, 10(8):707–710.

Meyers, A., M. Kosaka, and R. Grishman. 1998. A multilingual procedure for dictionary-based sentence alignment. In *Machine Translation and the Information Soup*, volume 1529 of *Lecture Notes in Computer Science*, pages 187–198. Springer.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sikes, R. 2007. Fuzzy matching in theory and practice. *MultiLingual*, 18(6):39–43.

Simard, M. 2003. Translation spotting for translation memories. In *Proceedings of the HLT-NAACL 2003, Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 65–72, Morristown, NJ, USA.

Somers, H., 2003. *Computers and translation: a translator's guide*, chapter Translation Memory Systems, pages 31–48. John Benjamins.

[4] http://opus.lingfil.uu.se/EMEA.php

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147, Genoa, Italy.

Tiedemann, J. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Borovets, Bulgaria.

Vogel, S., H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.

Véronis, J. and P. Langlais, 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora (Text, Speech and Language Technology)*, chapter Evaluation of Parallel Text Alignment Systems – The ARCADE Project, pages 369–388. Kluwer Academic Publishers.