Journal of Artificial Intelligence Research 53 (2015) 169-222 Submitted 11/14; published 06/15

Using Machine Translation to Provide Target-Language Edit Hints in Computer Aided Translation Based on Translation Memories

Miquel Esplà-Gomis MESPLA@DLSI.UA.ES
Felipe Sánchez-Martínez FSANCHEZ@DLSI.UA.ES
Mikel L. Forcada MLF@DLSI.UA.ES
Dept. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

Abstract

This paper explores the use of general-purpose machine translation (MT) in assisting the users of computer-aided translation (CAT) systems based on translation memory (TM) to identify the target words in the translation proposals that need to be changed (either replaced or removed) or kept unedited, a task we term as *word-keeping recommendation*. MT is used as a *black box* to align source and target sub-segments *on the fly* in the translation units (TUs) suggested to the user. Source-language (SL) and target-language (TL) segments in the matching TUs are segmented into overlapping sub-segments of variable length and machine-translated into the TL and the SL, respectively. The bilingual sub-segments obtained and the matching between the SL segment in the TU and the segment to be translated are employed to build the features that are then used by a binary classifier to determine the target words to be changed and those to be kept unedited. In this approach, MT results are never presented to the translator. Two approaches are presented in this work: one using a word-keeping recommendation system which can be trained on the TM used with the CAT system, and a more basic approach which does not require any training.

Experiments are conducted by simulating the translation of texts in several language pairs with corpora belonging to different domains and using three different MT systems. We compare the performance obtained to that of previous works that have used statistical word alignment for word-keeping recommendation, and show that the MT-based approaches presented in this paper are more accurate in most scenarios. In particular, our results confirm that the MT-based approaches are better than the alignment-based approach when using models trained on out-of-domain TMs. Additional experiments were also performed to check how dependent the MT-based recommender is on the language pair and MT system used for training. These experiments confirm a high degree of reusability of the recommendation models across various MT systems, but a low level of reusability across language pairs.

1. Introduction

Computer-aided translation (CAT) systems based on translation memory (TM) (Bowker, 2002; Somers, 2003) are the translation technology of choice for most professional translators, especially when translation tasks are repetitive and the effective recycling of previous translations is feasible. The reasons for this choice are the conceptual simplicity of *fuzzy-match scores* (FMS) (Sikes, 2007) and the ease with which they can be used to determine the usefulness of the translations proposed by the CAT system and to estimate the remain-

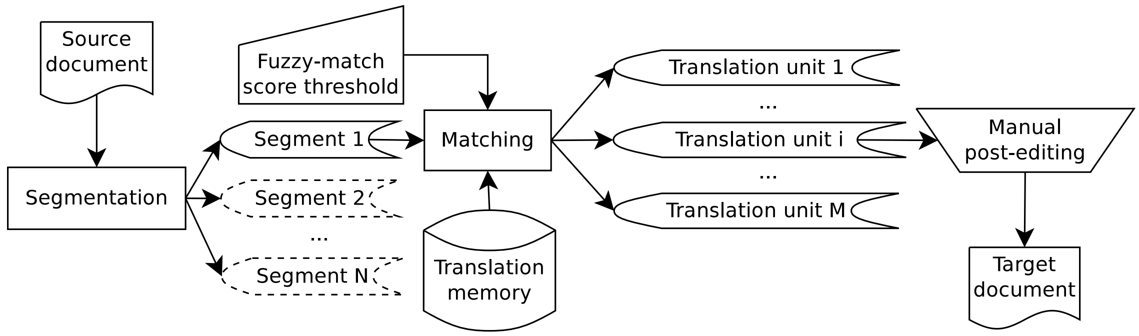

Figure 1: Procedure followed to translate a document using a TM-based CAT system.

ing effort needed to turn them into adequate translations. The FMS function measures the similarity between two text segments, usually by computing a variant of the word-based edit distance (Levenshtein, 1966), although the FMS of some proprietary tools are not publicly described.

When a TM-based CAT system is used to translate a new source document, the system first segments the document, and then, for each source segment S', provides the translator with the subset of translation units (TUs) (S,T) in the TM for which the FMS between S' and S is above a selected threshold Θ. The translator must then choose the TU (S,T) that best fits his or her needs and post-edit its target segment T to produce T', an adequate translation of S'. Figure 1 illustrates this procedure.

When showing the subset of matching TUs to the translator, most TM-based CAT systems highlight the words in S that differ from those in S' in order to ease the task of post-editing T. It is, however, up to the translator to identify the specific words in T that should be changed (either replaced or removed) in order to convert T into T', which is the problem that we deal with in this paper and term as *word-keeping recommendation*. Our experiments with professional translators show that a TM-based CAT system capable of word-keeping recommendation improves their productivity by up to 14% in the ideal case that all recommendations are indeed correct (see Appendix A for more details).

Word-keeping recommendation is related to *translation spotting* (Veronis & Langlais, 2000; Simard, 2003; Sánchez-Martínez, Carrasco, Martínez-Prieto, & Adiego, 2012), which consists of solving the problem of finding parallel sub-segments in parallel texts. Translation spotting is used, for example, by *bilingual concordancers* (Bourdaillet, Huet, Langlais, & Lapalme, 2010), types of tools which help a translator to retrieve occurrences of a sub-segment in a parallel corpus and its corresponding translation. Some examples of commercial bilingual concordancers are *Webitext*,[1] *Linguee*,[2] or *Reverso Context*.[3] Translation spotting is also particularly relevant for example-based machine translation (Somers, 1999), which uses this technique to build the sub-segmental TM used to translate new materials. MT quality estimation (de Gispert, Blackwood, Iglesias, & Byrne, 2013; Specia, Raj, & Turchi, 2010), also shares some features with this task: in both cases the objective is to discover whether

1. `http://www.webitext.com` [last visit: 15th May 2015]
2. `http://www.linguee.com` [last visit: 15th May 2015]
3. `http://context.reverso.net/translation/` [last visit: 15th May 2015]

a translation proposal T^4 is a valid translation for a given source language segment S'. The parallelisms become stronger in the case of word-level quality estimation (Ueffing & Ney, 2005; Bojar et al., 2014; Esplà-Gomis, Sánchez-Martínez, & Forcada, 2015), in which, as in word-keeping recommendation, every word of a proposal is analysed to decide whether or not it is likely to belong to the final translation. There are critical differences between the scenarios in which quality estimation and word-keeping recommendation operate: quality estimation detects words which should be changed in segments T which are likely to be inadequately written in TL, but are intended to be translations of S'; conversely, word-keeping recommendation is intended to work on segments T which are usually adequately written in TL, but they are not a translation of S' (unless an exact match between S and S' is found).

Esplà, Sánchez-Martínez, and Forcada (2011) have performed word-keeping recommendation by using statistical word-alignment models (Och & Ney, 2003) to align the source-language (SL) and target-language (TL) words of each TU in the TM. When a TU (S,T) is suggested to the translator, the pre-computed word alignments are then used to determine the target words to be changed or kept unedited. Analogously, Kranias and Samiotou (2004) align the words in each TU at different sub-segment levels by using, among other resources, a bilingual dictionary of words and phrases (Meyers, Kosaka, & Grishman, 1998), suffix lists to deal with morphological variations, and a list of closed-class words and their categories (Ahrenberg, Andersson, & Merkel, 2000). The authors use these alignments to detect the words to be changed and then use MT to propose a translation for them. To the best of our knowledge, the specific details on how the Kranias and Samiotou method works have not been published. A patent published by Kuhn, Goutte, Isabelle, and Simard (2011) describes a similar method that is also based on statistical word alignment in order to detect the words to be changed in a translation proposal. Unfortunately, the patent does not provide a detailed description of the actual procedure used.

Esplà-Gomis, Sánchez-Martínez, and Forcada (2011) follow a different approach which does not necessitate the computation of word alignments. Instead, they make use of any available MT system as a source of bilingual information to compute a set of features that are used by a perceptron classifier to estimate the probability p_K of each target word being kept unedited. This is done by: obtaining the matching TUs in the TM by using FMS above a given threshold; segmenting the SL and TL segments in each of these TUs into overlapping sub-segments of variable length; machine translating these sub-segments into the TL and the SL, respectively, in order to learn sub-segment alignments; and using these sub-segment alignments and the matching between S and S' to build the features to be used by the classifier. The basic idea behind this method is that a word in T is likely to be kept unedited if it appears in the translation of sub-segments which are common to S and S', the segment to be translated. Finally, p_K is used for word-keeping recommendation by marking the words for which $p_K < \frac{1}{2}$ as "change", or otherwise as "keep".

Although the latter approach requires a training procedure to be run on a TM, Esplà-Gomis et al. (2011) show that, for the translation of Spanish texts into English, the model used by the perceptron classifier can be trained on a TM from a domain that is different from that of the actual TM to be used and the text to be translated. Furthermore, the

4. In the case of quality estimation, the segment T to be evaluated originates from MT, while in word-keeping recommendation, it originates from a TM-based CAT tool proposal.

results obtained by this system are similar to those obtained by Esplà et al. (2011), based on statistical word alignments, for models trained on texts from the same domain as the text translated, and much better for models trained on out-of-domain texts, as shown in Section 7.

In this paper we revisit the approach of Esplà-Gomis et al. (2011), and propose new feature sets that capture the information in the machine-translated sub-segments in a more successful way than the features therein. In addition, a more complex multilayer perceptron binary classifier is used in this work, which improves the results obtained with the simpler perceptron classifier proposed by Esplà et al. (2011). These improvements on binary classification are compared to the previous approach on a more exhaustive evaluation framework, including new domains and language pairs (see below). Finally, we introduce a new method for word-keeping recommendation that is also able to use any available MT system as a source of bilingual information, and does not require any training procedure to be run in advance. This training-free method uses the sub-segment pairs that match both S and T to compute the *alignment strength* (Esplà-Gomis, Sánchez-Martínez, & Forcada, 2012b) between the words in S and T. The alignment strength between two words s_k in S and t_j in T measures the amount of evidence that relates the two words by giving more weight to the evidence from shorter sub-segments, which involves a sharper picture of the relation between s_k and t_j. Alignment strengths are then used in a similar fashion to that of Esplà et al. (2011) to determine the words to be changed or kept unedited.

As mentioned above, the experiments performed in this work compare the two MT-based approaches (that which requires training and that which is training-free) to the alignment-based approach of Esplà et al. (2011) using ten different language pairs, TMs from three different domains, and three different MT systems. The experiments not only cover the ideal scenario, in which the trained recommendation models are tested under the same conditions —language pair, TM domain, and MT system— used for training, but also scenarios in which some conditions change in order to test the reusability of the models. Namely, experiments were carried out by using recommendation models trained on: a TM from a different domain, a different MT system, and a different language pair. The results obtained show that the MT-based approaches are superior to the alignment-based approach as regards accuracy in all the scenarios. These results additionally confirm that the MT-based approaches produce recommendation models that are more portable across TM domains than those based on word alignment. This also provides good reusability for different MT systems, but poor reusability when translating a different pair of languages to that used for training; in fact, the training-free MT-based method provides better results in this scenario.

The remainder of the paper is organised as follows. The following section reviews the previous works on the integration of TMs and MT. Section 3 reviews the statistical word-alignment-based approach defined by Esplà et al. (2011), which is used in this paper as a reference to compare the new methods presented. Section 4 tackles the problem of word-keeping recommendation by using a binary classifier and several sets of features based on the coverage of sub-segment pairs obtained by machine translating sub-segments of different sizes from the segments in a translation proposal, and the matching between the source-side of the proposal and the segment to be translated. Section 5 then shows how to use these bilingual sub-segments to compute the alignment strength between the SL and TL words in each TU, and how they can be used for word-keeping recommendation without training. Section 6 describes the experimental framework, while Section 7 presents and discusses the results obtained. The paper ends with some concluding remarks. Two appendices are included in this paper: one including experiments aimed at measuring the impact of word-keeping recommendation on the productivity of professional translators, and the other one reporting the results on a filtered-out data set to check their performance in an ideal setting.

2. Integration of Machine Translation and Translation Memories

The literature on this subject contains several approaches that combine the benefits of MT and TMs in ways that are different from those presented in this paper, and that go beyond the obvious *β-combination* scenario defined by Simard and Isabelle (2009), in which MT is used to translate a new segment when no matching TU above a fuzzy-match score threshold β is found in the TM.

Marcu (2001) integrates word-based statistical machine translation (SMT) with a sub-segmental TM. The method uses *IBM model 4* (Brown, Della Pietra, Della Pietra, & Mercer, 1993) word-based translation model to build a sub-segmental TM and learn word-level translation probabilities. This is done by training the the IBM model 4 on the TM used for translation. The source language (SL) segments and the target language (TL) segments in each translation unit (TU) in the TM are then aligned at the word level by using the Viterbi algorithm. Finally, the sub-segmental TM is built from parallel *phrases* in a very similar way to that which occurs in modern phrase-based statistical MT systems (Koehn, 2010): parallel phrases are identified as all those pairs of sub-segments for which all the words on the SL side are aligned with a word on the TL side or with NULL (unaligned), and vice versa. The translation process is then carried out in two stages: first, occurrences of SL phrases in the sub-segmental TM are translated using the corresponding TL phrase; second, words not covered are translated using the word-level translation model learned by the IBM model 4. A similar approach is proposed by Langlais and Simard (2002), who also use translation at sub-segment level. In this case, the segment to be translated is split into sub-segments, and an online bilingual concordancer is used to find their translations. The word-based SMT decoder by Nießen, Vogel, Ney, and Tillmann (1998) is then used to choose the best sub-segments and put them in the best order according to the model.

Biçici and Dymetman (2008) integrate a phrase-based SMT (PBSMT) (Koehn, 2010) system into a TM-based CAT system using discontinuous bilingual sub-segments. The PBSMT system is trained on the same TM, and when a new source segment S' is to be translated, the segments S and T in the best matching TU are used to bias the statistical translation of S' towards T. This is done by augmenting the translation table of the PBSMT system with bilingual sub-segments originating from the fuzzy match (S, T) in which the source part is a common sub-sequence of S and S', and the target part is a sub-sequence of T which has been detected to be aligned with its counterpart sub-sequence in S. Simard and Isabelle (2009) propose a similar approach in which a new feature function is introduced in the linear model combination of a PBSMT system to promote the use of the bilingual sub-segments originating from the best fuzzy match (S, T). Following a similar approach, Läubli, Fishel, Volk, and Weibel (2013) use the mixture-modeling technique (Foster & Kuhn, 2007) to learn a domain-adapted PBSMT system combining an in-domain TM and more general parallel corpora. It is worth noting that none of these three approaches guarantees that the PBSMT system will produce a translation containing the translation in T of the sub-segments that are common to S and S'. In contrast, Zhechev and van Genabith (2010) and Koehn and Senellart (2010), who also use a PBSMT system, do guarantee that the sub-segments of T that have been detected to be aligned with the sub-segments in S matched by S' appear in the translation.

Example-based machine translation (EBMT) (Somers, 1999) has also frequently been used to take advantage of TMs at the sub-segment level (Simard & Langlais, 2001). EBMT systems are based on partial matches from TMs, as in the case of TM CAT tools. In this case, the matching TUs are aligned to detect sub-segment pairs that can be reused for translation. These sub-segment pairs are then combined to produce the most suitable translation for S'. For instance, the commercial TM-based CAT tool *Déjà Vu*[5] integrates example-based MT in order to suggest candidate translations in those cases in which an exact match is not found, but partial matches are available (Garcia, 2005; Lagoudaki, 2008). The example-based-inspired MT system is used to propose a translation by putting together sub-segments of the partial matchings available. Unfortunately, we have been unable to find further details on how this method works.[6] Approaches that combine several MT systems are also available. For example, Gough, Way, and Hearne (2002) use several online MT systems to enlarge the example database of an EBMT system. The authors claim that this permits a better exploitation of the parallel information in the TM for new translations.

Our approach differs from those described above in two ways. First, the aforementioned approaches use the TM to improve the results of MT, or use MT to translate sub-segments of the TUs, while the MT-based approaches presented in this paper use MT to improve the experience of using a TM-based CAT system *without actually showing any machine translated material to the translator*. Second, the approaches above, with the sole exception of that by Gough et al. (2002), focus on a specific MT system or family of MT systems (namely, SMT and EBMT), whereas our MT-based approaches use MT as a black box, and are therefore able to use one or more MT systems at once. In addition, as our MT-based approaches do not need to have access to the inner workings of the MT systems, they are capable of using MT systems that are available on-line (thus avoiding the need for local installation) or even any other source of bilingual information such as dictionaries, glossaries, terminology databases, or sub-segment pairs from bilingual concordancers.

Some works cited in this section, such as those by Zhechev and van Genabith (2010) and Koehn and Senellart (2010) use PBSMT, but they may easily be extended in order to use a different MT system. Their approaches share some similarities with ours: they also try to detect word or phrase alignments as a source of information to find out which parts of a translation proposal should be kept unedited. The main difference between our approach and those by Zhechev and van Genabith and Koehn and Senellart is that they use MT to produce a final translation for the segment to be translated, which comes closer to MT than to TM-based CAT. One of the aims of our approach is to minimally disturb the way translators work with the TM-based CAT system by keeping the translation proposals as found in the TM.

5. `http://www.atril.com` [last visit: 15th May 2015]
6. This is usually called "advanced leveraging" (Garcia, 2012).

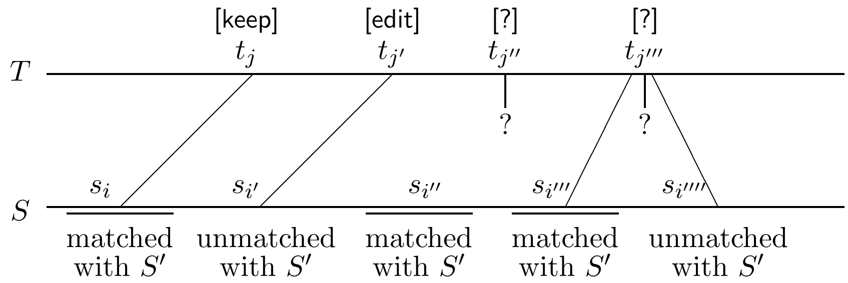

Figure 2: Example of the possible word-alignments that can be obtained for a pair of segments (S, T). Target word t_j may have to remain unedited because it is aligned with source word s_i which is in the part of S that matches S'. Target word $t_{j'}$ may have to be changed because it is aligned with source word $s_{i'}$ which is in the part of S that does not match S'. As target word $t_{j''}$ is not aligned with any source word in S, there is no evidence that could be used to make a recommendation for it. The case of word $t_{j'''}$ is special, since it is aligned with two words, one matching S' and the other not matching S', and a straightforward recommendation cannot be provided.

3. Word-Keeping Recommendation Based on Statistical Word Alignment

This section reviews the first approach for word-keeping recommendation, which was introduced by Esplà et al. (2011), who used statistical word alignment to detect the words to be kept or edited in a translation proposal. Given a segment S' to be translated and a TU (S, T) proposed by the TM-based CAT system, this method first computes the matching between S and S' and aligns the words in S and T by using the word-based statistical translation models implemented in GIZA++ (Och & Ney, 2003). Alignments are then used as follows: let t_j be the word in the j-th position of T which is aligned with a word s_i, the word in the i-th position of S. If s_i is part of the matching between S and S', this indicates that t_j might be part of the translation of S' and that it should therefore remain unedited, as occurs with the word t_j in Figure 2. Conversely, if s_i is not part of the match between S and S', this indicates that t_j might not be the translation of any of the words in S' and it should be edited, as occurs with word $t_{j'}$ in Figure 2. More complex situations in which a TL word is aligned with more than one SL word are tackled by following a voting scheme, as will be explained below. The main limitation of this approach is that, when a word t_j is unaligned, as occurs with the word $t_{j''}$ in Figure 2, there is no evidence that could be used to make a recommendation for it. Although it might be possible to decide on unaligned words by, for example, using the aligned words surrounding them, a wrong recommendation could be worse for the translator than not making any recommendation at all. The idea behind this claim is that a wrong keep recommendation may lead to a wrong translation, which would clearly be undesirable.

In order to determine whether the word t_j in the target proposal T should be changed or kept unedited, the fraction of words aligned with t_j which are common to both S and S'

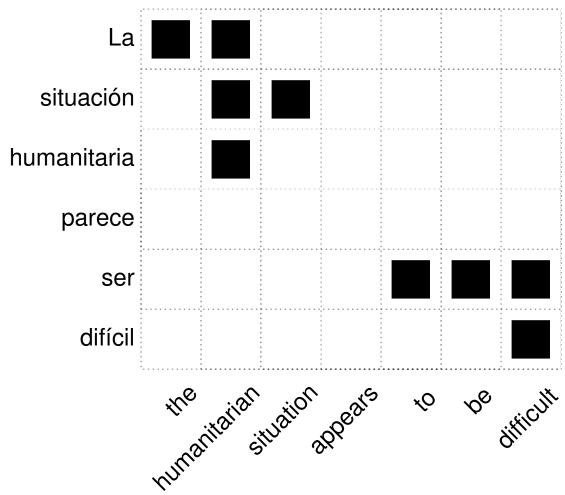

Figure 3: Word alignments for the TU *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult")*.

are computed:

$$f_K(t_j, S', S, T) = \frac{\displaystyle\sum_{s_i \in \mathrm{aligned}(t_j)} \mathrm{matched}(s_i)}{|\mathrm{aligned}(t_j)|}$$

where aligned(t_j) is the set of source words in S which are aligned with target word t_j in T, and matched(s_i) equals 1 if the word s_i is part of the match between S and S', the segment to be translated, and 0 otherwise. Function matched(x) is based on the optimal edit path,[7] obtained as a result of the word-based edit distance (Levenshtein, 1966) between S and S'. The fraction $f_K(t_j, S', S, T)$ may be interpreted as the likelihood of having to keep word t_j unedited. As mentioned above, t_j may be aligned with several words in S, some of which may be common to S and S' while others may not, as occurs with the word $t_{j'''}$ in Figure 2. Esplà et al. (2011) propose two possible heuristics to deal with this:

- *unanimity*: for a word t_j, a recommendation is made if it is aligned only with matched words ($f_K(\cdot) = 1$), or only with unmatched words ($f_K(\cdot) = 0$) in S, while no recommendation is made otherwise; and

- *majority*: this heuristic uses a voting scheme, in which if t_j is aligned with more matched words than unmatched words ($f_K(\cdot) > \frac{1}{2}$), a recommendation is made that it should be kept, and vice versa. Only if t_j is aligned with the same number of matched and unmatched words ($f_K(\cdot) = \frac{1}{2}$) no recommendation is made.

Let us suppose that the TU $(S, T) =$ *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult")* is proposed in order to translate a new

7. It may occur that more than one optimal path is available to align two segments S and S'. In this case, one of them is chosen arbitrarily.

segment S' = *"la situación política parece ser difícil"*, and that the word-alignment for (S, T) is that depicted in Figure 3. The words *the*, *situation*, *to* and *be* would be marked to be kept, since they are aligned with a single word which is part of the matching between S and S', which is compatible with a possible translation T'= *"the political situation appears to be difficult"*. The word *difficult* would also be marked to be kept, since, even though it is aligned with two words, both are part of the matching between S and S'. However, the evidence for the word *humanitarian* is ambiguous; it is aligned with the words *la* and *situación*, which are part of the matching, but also with *humanitaria* which is not. If the unanimity criterion were to be used, no recommendation would be made for it, while the use of the majority criterion would result in a keeping recommendation. Finally, no recommendation would be made for the word *appears*, since it is not aligned with any word.

The main disadvantage of this approach is that it requires a word-alignment model to be trained directly on the TM to be used for translation in order to maximise the coverage, which means re-training the alignment model every time the TM is updated with new TUs. It may also occur that the TM is not sufficiently large to be able to obtain recommendations with an acceptable quality, signifying that it is necessary to use external parallel corpora in order to train these models. Incremental training (Gao, Lewis, Quirk, & Hwang, 2011) or online training (Bertoldi, Farajian, & Federico, 2009) of statistical word alignment models could be a means to reduce the training time when a TM is modified, or even to adapt a general alignment models to more specific domains, thus improving the coverage. In this case, incremental training would be useful as regards adapting existing models to a new TM, while the on-line training would allow the models to be updated after a new TU has been added to a TM. However, this paper focuses on using machine translation as a source of bilingual information for word-keeping recommendation: we therefore keep the original word-alignment-based approach as described by Esplà et al. (2011) and only use it as a reference when comparing the new approaches proposed here.

4. Word-Keeping Recommendation as Binary Classification

In this work we tackle the problem of word-keeping recommendation as a binary classification problem. For a new segment S' and the TU (S, T) suggested to the translator by the TM-based CAT system, a set of features are computed for each word t_j in T, and a binary classifier is then used to determine whether t_j should be kept unedited or changed (either replaced or deleted). Henceforth, we shall refer to this approach as the *trained MT-based recommender*, to differentiate it from the *training-free MT-based recommender* presented in Section 5.

The features we use are based on the assumption that MT, or any other source of bilingual information, can provide evidence as to whether each word t_j in T should be changed or kept unedited. Let σ be a sub-segment of S from one of the matching TUs (S, T), which is *related by MT* to a sub-segment τ of T. By *related by MT* we mean that machine translating σ leads to τ, or vice versa. We hypothesise that:

- words in σ matching the new segment to be translated S' provide evidence that the words in τ should be kept unedited ("keeping evidence"); and

(“la”,“the”) ***[“keeping evidence”]****,*
(“situación”,“situation”) ***[“keeping evidence”]****,*
(“humanitaria”, “humanitarian”) ***[“changing evidence”]****,*
(“ser”,“be”) ***[“keeping evidence”]****,*
(“ser”,“to be”) ***[“keeping evidence”]****,*
(“difícil”,”difficult”) ***[“keeping evidence”]****,*
(“situación humanitaria”,“humanitarian situation”),
(“ser difícil”, “be difficult”) ***[“keeping evidence”]****,* and
(“la situación humanitaria”, “the humanitarian situation”).

Figure 4: Example of the collection M of overlapping machine translated pairs of sub-segments for (S, T) = (“la situación humanitaria parece ser difícil”, “the humanitarian situation appears to be difficult”). Sub-segment pairs (σ, τ) with σ matching S and τ matching T are highlighted in bold type.

- words in σ not matching the new segment to be translated S' provide evidence that the words in τ should be changed (“changing evidence”).

To continue with the example proposed in Section 3, in which: (S, T) = *(“la situación humanitaria parece ser difícil”, “the humanitarian situation appears to be difficult”),* and S'= *“la situación política parece ser difícil”,* we segment S and T into all possible overlapping sub-segments and translate them with an MT system to obtain the collection M of sub-segment pairs (σ, τ) matching (S, T) and shown in Figure 4. Some sub-segment pairs, such as *(“parece”, “appear”)*, are not included in that list because the translations of the sub-segment on one side do not match their equivalents on the other side. For example, *parece* is translated into English as *seems*, while *appear* is translated into Spanish as *aparecer*. Those pairs (σ, τ) in that list for which all the words in σ match S' provide strong evidence that the words in the corresponding target part should be kept unedited. In this example, these words are *the*, *situation*, *be* and *difficult*, which are compatible with a possible translation T'=*“the political situation appears to be difficult”*. Conversely, the pairs (σ, τ) for which all the words in σ do not match S' provide strong evidence that the words in the target part should be changed. In this case, this only occurs for the word *humanitarian*. On the other hand, there is one word about which no evidence can be obtained (*appears*) because it is not matched by the MT system. In this case, it is not possible to provide the translator with any recommendations as occurs, for analogous reasons, with the alignment-based approach described in Section 3. Note that the pair (σ, τ) =*(“situación humanitaria”, “humanitarian situation”)* contains a source word (*situación*) which matches S' and another (*humanitaria*) which does not match S'. Dealing with this ambiguous evidence, along with combining the evidence from different (σ, τ) (which may be contradictory) leads to an additional problem. In order to deal with ambiguous evidence, we define three feature sets that model and combine “keeping” and “changing” evidence, and which will be described in Sections 4.1, 4.2, and 4.3.

It is worth noting that some pre-processing methods could be used in order to, hopefully, exploit the evidence from bilingual sources of information more efficiently, such as stemming/lemmatisation, morphological analysis, or even the integration of syntactic features such as those proposed by Ma, He, Way, and van Genabith (2011). However, the objective of this approach is to avoid any complex processing in order to obtain fast recommendations when translating texts between any pair of languages, in any domain, and only re-using already available sources of bilingual information, such as the numerous MT systems available on the Internet.

4.1 Features Based on Matching/Mismatching Sub-segments with Unconstrained Length Relations [MM-U]

This feature set was proposed by Esplà-Gomis et al. (2011) and is used as a reference for the remaining feature sets proposed in this work. This feature set considers, for a given (σ, τ) pair of segments, that:

- if σ is a common sub-segment of both the new segment to be translated S' and the source segment S, then it is likely that the words in τ will not have to be changed ("keeping evidence");
- if σ is a sub-segment of S but not of S', then it is likely that the words in τ will have to be changed ("changing evidence").

As can be seen, this is a rather conservative criterion which discards the information from matching words in a partially matching sub-segment σ between S' and S,[8] and will probably be capable of providing high accuracy when recommending that a word should be kept. A more flexible approach is presented in Section 4.2.

Based on the proposed rationale, four sets of features are computed: two sets of *keeping features*, which provide information about the chances of keeping t_j, and two sets of *changing features*, which provide information about the chances of changing t_j. Given the maximum sub-segment length L, the keeping feature set K_{m*} is defined for the word t_j and for every value of $m \in [1, L]$ as follows:

$$K_{m*}(j, S', S, T) = \frac{\text{tcover}(j, \text{seg}_m(S) \cap \text{seg}_m(S'), \text{seg}_*(T), M)}{\text{tcover}(j, \text{seg}_m(S), \text{seg}_*(T), M)},$$

where $\text{seg}_m(X)$ represents the set of all possible m-word sub-segments of segment X, $\text{seg}_*(X)$ is similar to $\text{seg}_m(X)$ but without length constraints, and $\text{tcover}(j, \mathcal{S}, \mathcal{T}, M)$ is defined as:

$$\text{tcover}(j, \mathcal{S}, \mathcal{T}, M) = |\{\tau \in \mathcal{T} : \exists \sigma \in \mathcal{S} \wedge (\sigma, \tau) \in M \wedge j \in \text{span}(\tau, T)\}|,$$

where $\mathcal{S} \subseteq \text{seg}_*(S)$, $\mathcal{T} \subseteq \text{seg}_*(T)$, and function $\text{span}(\tau, T)$ returns the set of word positions spanned by the sub-segment τ in the segment T.[9] Function $\text{tcover}(j, \mathcal{S}, \mathcal{T}, M)$ therefore computes the number of target sub-segments $\tau \in \mathcal{T}$ containing the word t_j that are related by MT to a sub-segment $\sigma \in \mathcal{S}$.

8. For example, a sub-segment of 5 words in which 4 of them are matched and only one is unmatched would be considered as "changing evidence".
9. Note that a sub-segment τ may be found more than once in segment T: function $\text{span}(\tau, T)$ returns all the possible positions spanned.

Similarly to K_{m*}, K_{*n} is computed by using only target sub-segments τ of length n:

$$K_{*n}(j, S', S, T) = \frac{\text{tcover}(j, \text{seg}_*(S) \cap \text{seg}_*(S'), \text{seg}_n(T), M)}{\text{tcover}(j, \text{seg}_*(S), \text{seg}_n(T), M)}.$$

Analogously, changing feature sets C_{m*} and C_{*n} are defined as:

$$C_{m*}(j, S', S, T) = \frac{\text{tcover}(j, \text{seg}_m(S) - \text{seg}_m(S'), \text{seg}_*(T), M)}{\text{tcover}(j, \text{seg}_m(S), \text{seg}_*(T), M)},$$

$$C_{*n}(j, S', S, T) = \frac{\text{tcover}(j, \text{seg}_*(S) - \text{seg}_*(S'), \text{seg}_n(T), M)}{\text{tcover}(j, \text{seg}_*(S), \text{seg}_n(T), M)}.$$

In the case of all four features, when both the numerator and the denominator happen to be zero because no pair (σ, τ) covers t_j, the value of the feature is set to $\frac{1}{2}$.

These four features are computed for every value of $1 \leq m \leq L$ and $1 \leq n \leq L$, where L is the maximum sub-segment length used, resulting in $4L$ features. All these features take values in $[0, 1]$ and may have a probabilistic interpretation, in which $\frac{1}{2}$ means "don't know". This feature set will from here on be termed as *MM-U* features. A similar collection of features was tried that constrained the length of both σ (m) and τ (n). However, the results confirmed that no improvement was obtained by adding it to the feature set.

For the running example, we show the feature set that could be computed at word *be*, the sixth word in T (t_6). Please recall that $(S, T) =$ *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult").* Using the collection of translated pairs of overlapping sub-segments shown in Figure 4, there are three sub-segment pairs (σ, τ) in M that cover the word *be*:

$$\boldsymbol{M} = \{(\text{“}\textbf{ser}\text{”}, \text{“}\textbf{be}\text{”}), (\text{“}\textbf{ser}\text{”}, \text{“}\textbf{to be}\text{”}), (\text{“}\textbf{ser difícil}\text{”}, \text{“}\textbf{be difficult}\text{”})\}.$$

These pairs (σ, τ) only contain sub-segments σ with $m \in [1, 2]$. The value of the function tcover is:

$$\text{tcover}(6, \text{seg}_1(S), \text{seg}_*(T), M) = |\{\text{“be”}, \text{“to be”}\}| = 2$$

$$\text{tcover}(6, \text{seg}_2(S), \text{seg}_*(T), M) = |\{\text{“be difficult”}\}| = 1$$

for both values of m. In addition, the value of tcover for all the sub-segments σ that match S' is:

$$\text{tcover}(6, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M) = |\{\text{“be”}, \text{“to be”}\}| = 2$$

$$\text{tcover}(6, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M) = |\{\text{“be difficult”}\}| = 1$$

The value of the corresponding features is therefore:

$$K_{1*}(6, S', S, T) = \frac{\text{tcover}(6, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M)}{\text{tcover}(6, \text{seg}_1(S), \text{seg}_*(T), M)} = \frac{2}{2} = 1$$

$$K_{2*}(6, S', S, T) = \frac{\text{tcover}(6, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M)}{\text{tcover}(6, \text{seg}_2(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

Features $K_{*1}(6, S', S, T)$ and $K_{*2}(6, S', S, T)$ can be computed analogously. This case is rather simple, since all the evidence available for this word indicates that it should be kept.

However, for the word *situation* (t_3), both "keeping" and "changing" evidence coexist in the set M of translated sub-segments pairs:

$$\boldsymbol{M} = \{(\text{“\textbf{situación}”}, \text{“\textbf{situation}”}), (\text{“situación humanitaria”}, \text{“humanitarian situation”}), \\ (\text{“la situación humanitaria”}, \text{“the humanitarian situation”})\}$$

In this case, τ sub-segments take lengths $m \in [1, 3]$, which produces the following values for $\text{tcover}(\cdot)$:

$$\text{tcover}(3, \text{seg}_1(S), \text{seg}_*(T), M) = |\{\text{“situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_2(S), \text{seg}_*(T), M) = |\{\text{“humanitarian situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_3(S), \text{seg}_*(T), M) = |\{\text{“the humanitarian situation”}\}| = 1$$

However, in this case, not all of them match S':

$$\text{tcover}(3, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M) = |\{\text{“situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M) = |\emptyset| = 0$$

$$\text{tcover}(3, \text{seg}_3(S) \cap \text{seg}_3(S'), \text{seg}_*(T), M) = |\emptyset| = 0$$

This allows us to compute the following keeping features:

$$K_{1*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_1(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

$$K_{2*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_2(S), \text{seg}_*(T), M)} = \frac{0}{1} = 0$$

$$K_{3*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_3(S) \cap \text{seg}_3(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_3(S), \text{seg}_*(T), M)} = \frac{0}{1} = 0$$

Analogously, for the changing features, we have:

$$\text{tcover}(3, \text{seg}_1(S) - \text{seg}_1(S'), \text{seg}_*(T), M) = |\emptyset| = 0$$

$$\text{tcover}(3, \text{seg}_2(S) - \text{seg}_2(S'), \text{seg}_*(T), M) = |\{\text{“humanitarian situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_3(S) - \text{seg}_3(S'), \text{seg}_*(T), M) = |\{\text{“the humanitarian situation”}\}| = 1$$

which allow us to obtain the following features:

$$C_{1*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_1(S) - \text{seg}_1(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_1(S), \text{seg}_*(T), M)} = \frac{0}{1} = 0$$

$$C_{2*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_2(S) - \text{seg}_2(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_2(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

$$C_{3*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_3(S) - \text{seg}_3(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_3(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

In this case, the ambiguity in the features will be managed by the binary classifier, which will determine the corresponding weights during training.

4.2 Features Based on Partially Matching Sub-segments with Constrained Length Relations [PM-C]

This feature set is slightly different from the previous one as regards the way in which the evidence from the pairs of sub-segments $(\sigma, \tau) \in M$ is used. In this case, the features represent the fraction of words in S that match S' to which a given word t_j in T is related by means of sub-segment pairs (σ, τ). It is worth noting that in the previous feature set, the matching of sub-segment pairs (σ, τ) was evaluated for the whole sub-segment σ. However, in this new feature set, both the keeping and the changing features are computed using the matched/unmatched words in σ. The objective of this feature set is to use the positive evidence from partially matching sub-segments σ more efficiently. The following equation defines the new keeping feature K_{mn}^W:

$$K_{mn}^W(j, S', S, T) = \sum_{k=1}^{|S|} \mathrm{stcover}(j, k, \mathrm{seg}_m(S), \mathrm{seg}_n(T), M) \times \mathrm{match}(k, S', S)$$

where j is the position of t_j in T, k is the position of s_k in S, $\mathrm{match}(k, S', S)$ is 1 if s_k is part of the match between S and S', and 0 otherwise,[10] and function $\mathrm{stcover}(j, k, \mathcal{S}, \mathcal{T}, M)$ is defined as:

$$\mathrm{stcover}(j, k, \mathcal{S}, \mathcal{T}, M) = |\{(\sigma, \tau) \in M : \tau \in \mathcal{T} \land \sigma \in \mathcal{S} \land j \in \mathrm{span}(\tau, T) \land k \in \mathrm{span}(\sigma, S)\}|$$

Similarly, we define the *changing feature* C_{mn}^W as:

$$C_{mn}^W(j, S', S, T) = \sum_{k=1}^{|S|} \mathrm{stcover}(j, k, \mathrm{seg}_m(S), \mathrm{seg}_n(T), M) \times (1 - \mathrm{match}(k, S', S)).$$

Function $\mathrm{stcover}(j, k, \mathcal{S}, \mathcal{T}, M)$ differs from $\mathrm{tcover}(j, \mathcal{S}, \mathcal{T}, M)$ in that, for a given pair (σ, τ), the former takes into account both σ and τ while the latter only takes into account τ. This makes K_{mn}^W and C_{mn}^W complementary, whereas K_{mn} and C_{mn} are not. K_{mn}^W and C_{mn}^W may be combined in a single normalised feature that we term as KC_{mn}^W:

$$\mathrm{KC}_{mn}^W(j, S', S, T) = \frac{\displaystyle\sum_{k=1}^{|S|} \mathrm{stcover}(j, k, \mathrm{seg}_m(S), \mathrm{seg}_n(T), M) \times \mathrm{match}(k, S', S)}{\displaystyle\sum_{k=1}^{|S|} \mathrm{stcover}(j, k, \mathrm{seg}_m(S), \mathrm{seg}_n(T), M)}, \tag{1}$$

As in the feature set described in Section 4.1, KC_{mn}^W takes values in $[0, 1]$, and, as in that case, when no evidence is found for t_j, the value of the corresponding feature is set to $\frac{1}{2}$. This feature set results in L^2 features and will be referred to as *PM-C*.

10. The function $\mathrm{match}(k, S', S)$ is based on the optimal edit path obtained as a result of the word-based edit distance (Levenshtein, 1966) between S' and S. Although this is not frequent, it may occur that more than one optimal paths are available: in this case, one of them is chosen arbitrarily.

For the running example, we compute the PM-C features for the word *situation*, as occurred in Section 4.1. As in the previous example, we use the collection of translated sub-segments in Figure 4. The set of sub-segments pairs covering the word *situation* is:

$$\boldsymbol{M} = \boldsymbol{\{}(\text{“}\textbf{situación}\text{”}, \text{“}\textbf{situation}\text{”}), (\text{“situación humanitaria”}, \text{“humanitarian situation”}), \\ (\text{“la situación humanitaria”}, \text{“the humanitarian situation”})\boldsymbol{\}}$$

and the features $KC^W_{1,1}(3, S', S, T)$, $KC^W_{2,2}(3, S', S, T)$, and $KC^W_{3,3}(3, S', S, T)$ can be computed from them. As can be seen stcover$(\cdot)$ happens to be different to zero only for $k = 1$:

$$\text{stcover}(3, 1, \text{seg}_1(S), \text{seg}_1(T), M) = |\{(\text{“}\textbf{situación}\text{”}, \text{“situation”})\}| = 1$$

In this case, we see that *situación* (s_2) is related to *situation* through the sub-segment pair (σ, τ)=(“situación”, “situation”). In this case, σ completely matches S', and we therefore have that:

$$KC^W_{1,1}(3, S', S, T) = \frac{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_1(S), \text{seg}_1(T), M) \times \text{match}(k, S', S)}{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_1(S), \text{seg}_1(T), M)} = \frac{1}{1} = 1$$

The case of $KC^W_{2,2}(3, S', S, T)$ is slightly more complex. Here, stcover$(\cdot)$ happens to be different to zero only for $k \in [1, 2]$:

$$\text{stcover}(3, 1, \text{seg}_2(S), \text{seg}_2(T), M) = \\ |\{(\text{“}\textbf{situación}\text{ humanitaria”}, \text{“humanitarian situation”})\}| = 1$$

$$\text{stcover}(3, 2, \text{seg}_2(S), \text{seg}_2(T), M) = \\ |\{(\text{“situación }\textbf{humanitaria}\text{”}, \text{“humanitarian situation”})\}| = 1$$

As will be observed, the words *situación* and *humanitaria* are related to *situation* through the same pair (σ, τ)=(“situación humanitaria”,“humanitarian situation”). Here, only one of the two words matches S', hence:

$$KC^W_{2,2}(3, S', S, T) = \frac{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_2(S), \text{seg}_2(T), M) \times \text{match}(k, S', S)}{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_2(S), \text{seg}_2(T), M)} = \frac{1}{2} = 0.5$$

Finally we have that, for $KC^W_{3,3}(3, S', S, T)$, stcover$(\cdot)$ happens to be different to zero for $k \in [1, 3]$:

$$\text{stcover}(3, 1, \text{seg}_3(S), \text{seg}_3(T), M) = \\ |\{(\text{“}\textbf{la}\text{ situación humanitaria”}, \text{“the humanitarian situation”})\}| = 1$$

$$\text{stcover}(3, 2, \text{seg}_3(S), \text{seg}_3(T), M) = |\{(\text{“la } \textbf{situación} \text{ humanitaria”}, \text{“the humanitarian situation”})\}| = 1$$

$$\text{stcover}(3, 3, \text{seg}_3(S), \text{seg}_3(T), M) = |\{(\text{“la situación } \textbf{humanitaria}\text{”}, \text{“the humanitarian situation”})\}| = 1$$

This time, three words are related to *situation*, all of them through the same sub-segment pair (σ, τ)=(“la situación humanitaria”,“the humanitarian situation”). In this case, *la* and *situación* match S', while *humanitaria* does not. The resulting feature is therefore:

$$KC^W_{3,3}(3, S', S, T) = \frac{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_3(S), \text{seg}_3(T), M) \times \text{match}(k, S', S)}{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_3(S), \text{seg}_3(T), M)} = \frac{2}{3} \simeq 0.67$$

Note that this feature collection constrains the length of σ and τ at the same time. This configuration was also tried in the previous feature set (MM-U), but no improvements were obtained as compared to constraining the lengths of σ and τ separately. With this feature set both possibilities were also tried, but here constraining the length of σ and τ at the same time proved to lead to better results.

4.3 Features Combining Partially Matching Sub-segments with Constrained Length Relations and Information about Coverage [PM-C+C]

Features KC^W_{mn} above may hide the amount of “keeping/changing evidence”, since they only take into account the fraction of “keeping evidence” from the total amount of evidence.[11] To deal with this, we propose that the feature set defined in Section 4.2 be combined with a new feature E_{mn}:

$$E_{mn}(j, S, T) = |\{(\sigma, \tau) \in M : \sigma \in \text{seg}_m(S) \wedge \tau \in \text{seg}_n(T) \wedge j \in \text{span}(\tau, T)\}| \qquad (2)$$

This feature counts the number of sub-segment pairs (σ, τ) covering word t_j, thus providing a measure of the amount of evidence supporting the value of feature KC^W_{mn}. We propose a new feature set, with $2L^2$ features, using both KC^W_{mn} and $E_{mn}(j, S', S, T)$, which will be referred to as *PM-C+C*. A similar feature set was tried, in which E_{mn} was normalised by dividing it by the maximum number of pairs (σ, τ) that could have covered t_j $(m * n)$. However, this set of features did not show any improvement and was therefore discarded.

For the running example, the pairs (σ, τ) in M that cover the word *be* (t_6) are:

$$\boldsymbol{M} = \{(\text{“}\textbf{ser}\text{”}, \text{“}\textbf{be}\text{”}), (\text{“}\textbf{ser}\text{”}, \text{“}\textbf{to be}\text{”}), (\text{“}\textbf{ser difícil}\text{”}, \text{“}\textbf{be difficult}\text{”})\}.$$

Therefore, the features E_{mn} that are different to zero for the word *be* are:

$$E_{1,1}(6, S, T) = |\{(\text{“ser”}, \text{“be”})\}| = 1$$

11. For example, it would be the same to have 1 “keeping evidence” out of 1 evidence and 5 keeping evidences out of 5 evidences; however, the second case should be considered to be more reliable, since more evidence confirms the keeping recommendation.

$$E_{1,2}(6, S, T) = |\{(\text{“ser”}, \text{“to be”})\}| = 1$$

$$E_{2,2}(6, S, T) = |\{(\text{“ser difícil”}, \text{“be difficult”})\}| = 1$$

Given that a single pair (σ, τ) covers the word *be* with the same m and n, all these features are set to 1. However, if the evidence (σ, τ)=*(“ser”, “be difficult”)* could be used, the value of $E_{1,2}$ would become higher:

$$E_{1,2}(6, S, T) = |\{(\text{“ser”}, \text{“to be”}), (\text{“ser”}, \text{“be difficult”})\}| = 2$$

5. Word-Keeping Recommendation Based on MT Alignment Strengths

The collection M of sub-segment pairs (σ, τ) which are related by MT can be used for word-keeping recommendation directly, i.e., without having to run any training procedure. We propose to do this by using a *training-free* MT-based recommender based on the *alignment strengths* described by Esplà-Gomis, Sánchez-Martínez, and Forcada (2012a) and Esplà-Gomis et al. (2012b). This metric determines the relatedness or association strength between the j-th word in T and the k-th word in S and is defined as:

$$A(j, k, S, T) = \sum_{m=1}^{L} \sum_{n=1}^{L} \frac{\text{stcover}(j, k, \text{seg}_m(S), \text{seg}_n(T), M)}{m \times n}$$

The alignment strength is based on the idea that matched sub-segment pairs apply *pressure* to the words, signifying that the larger the surface covered by a sub-segment pair, the lower the pressure applied to each individual word. Figure 5 shows how the words in a TU are covered by the bilingual sub-segments in M (left), and the result of computing the alignment strengths (right).

In order to perform a word-keeping recommendation using the alignment strengths, we define function $G(j, S', S, T)$ which computes the fraction of the alignment strength that relates a word t_j to those words s_k which are part of the matching between S' and S, over the sum of the alignment strength for all the words in S:

$$G(j, S', S, T) = \frac{\sum_{k=1}^{|S|} A(j, k, S, T) \times \text{match}(k, S', S)}{\sum_{k=1}^{|S|} A(j, k, S, T)} \tag{3}$$

Word keeping recommendation is then performed in the following simple manner: if $G(j, S', S, T) \leq \frac{1}{2}$, then word t_j is marked to be changed, otherwise to be kept. If there is no evidence (σ, τ) with τ spanning t_j, then no recommendation is provided for t_j.

It is worth noting that $A(j, k, S, T)$ is similar to a particular linear combination of the PM-C feature set described in Section 4.2, in which the weight of each feature is directly set to $\frac{1}{mn}$, rather than being chosen by optimising the recommendation accuracy in a training set. The results shown in Section 6 prove that this method is less accurate than the trained MT-based approach, but still provides reasonably good results.

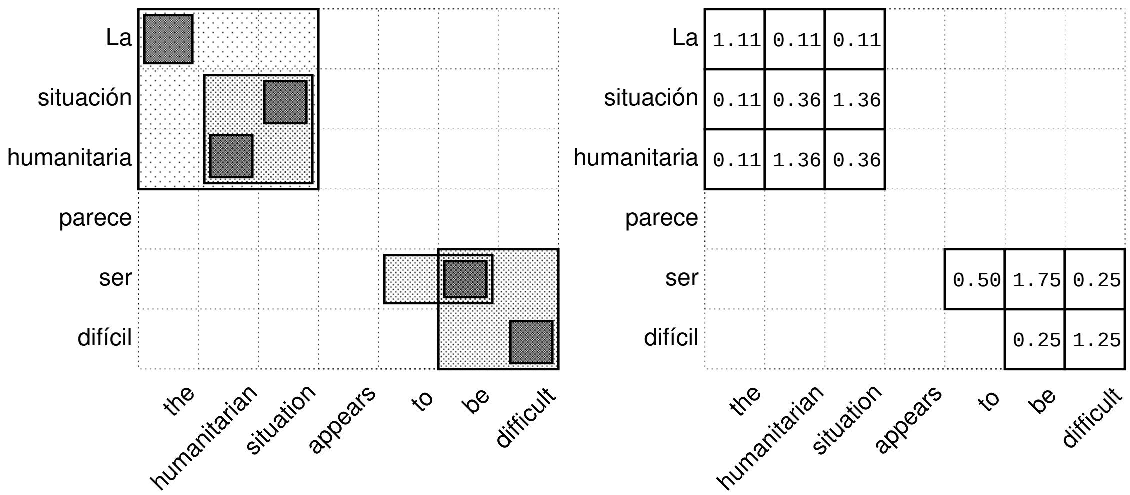

Figure 5: Sub-segment pairs covering the words in the TU *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult")* (left), and the alignment strengths obtained from them (right). The weight of each sub-segment pair is taken to be 1 and is divided by the surface it covers to compute the "pressure" exerted on each individual word.

For the running example, we use the scores shown in Figure 5; as can be seen, the word *situation* is related to three words: *La*, with a score of 0.11, *situación*, with a score of 1.36, and *humanitaria*, with a score of 0.36. The value of $G(3, S', S, T)$ is therefore:

$$G(3, S', S, T) = \frac{0.11 \times 1 + 1.36 \times 1 + 0.36 \times 0}{0.11 + 1.36 + 0.36} = \frac{1.47}{1.83} \simeq 0.8$$

In this case, $G(3, S', S, T) > \frac{1}{2}$, which means that the word *situation* must remain unedited. However, for *humanitarian*, which is related to the same words in Spanish, we have:

$$G(2, S', S, T) = \frac{0.11 \times 1 + 0.36 \times 1 + 1.36 \times 0}{0.11 + 0.36 + 1.36} = \frac{0.47}{1.83} \simeq 0.3$$

Since $G(2, S', S, T) \leq \frac{1}{2}$, the word *humanitarian* would be marked to be changed.

6. Experimental Settings

The experiments conducted consisted of simulating the translation of texts between several language pairs and text domains. The language pairs involved in the experiments were German–English (de→en), English–German (en→de), English–Spanish (en→es), Spanish–English (es→en), English–Finnish (en→fi), Finnish–English (fi→en), English–French (en→fr), French–English (fr→en), Spanish–French (es→fr), and French–Spanish (fr→es). Three thematic TMs were created for each language pair by extracting domain-specific TUs from the DGT-TM (Steinberger, Eisele, Klocek, Pilos, & Schlüter, 2012), a TM published by the *European Commission Directorate-General for Translation* (European Commission Directorate-General for Translation, 2009).

We compared the two MT-based approaches described in Sections 4 and 5 with a *naïve* baseline which is also based on a binary classifier, but using only the fuzzy-match score (FMS) between S and S' as a feature, (henceforth *FMS-only baseline*, see Section 6.5), and with the statistical word-alignment-based approach described in Section 3, in different scenarios. We first evaluated all the approaches in the optimal case, in which the models (either word-alignment models or classification models) are trained on the same TM used for translating, and, in the case of the MT-based approaches, employing the same MT system used for translation. This evaluation was then extended to evaluate:

- *the reusability across domains*: the word-alignment models and the classification models are trained on out-of-domain TMs;
- *the reusability across MT systems*: the models are trained on the same TM used for translation, but using a different MT system; and
- *the reusability across language pairs*: the models are trained on a TM from the same domain as that used for translation, but with a different language pair, and obviously, using a different MT system.

The reusability across MT systems can evidently be evaluated only in the case of MT-based approaches. As regards reusability across language pairs, on the one hand, the FMS-only baseline is language independent, while on the other, the statistical word-alignment models used by the alignment-based approach have to be trained on the same pair of languages as that used for translation.

This extensive evaluation will allow us to ascertain the degree of independence of the recommendation model with regard to the domain of the TM, the MT system, and the language pair used during training. This is a key point, since high independence of all or part of these variables would allow computer-aided translation (CAT) users to reuse existing feature weights obtained without having to run any training procedure when they change the domain of the texts to be translated, the MT system they use or even the languages they are working with. The case of domain independence is particularly relevant since it covers not only the problem of using a different TM, but also the case in which new TUs which were not seen during training are added to a TM.

With regard to the MT systems, we have used the statistical MT system by Google,[12] Power Translator (Depraetere, 2008) version 15,[13] and the free/open-source, shallow-transfer MT system Apertium (Forcada et al., 2011).[14] Unfortunately, not all the MT systems mentioned above were available for all the language pairs. Table 1 shows the MT system(s) available for each language pair included in the experiments.

Even though we used large data sets in a batch mode to obtain the results reported in this paper, we wanted to ensure that the MT-based approaches would be able to provide recommendations in real time for translation tasks. The main part of the computation time for the MT-based approaches is spent segmenting S and T and machine-translating the resulting sub-segments. In order to prove that this could be done in a real MT-based

12. `http://translate.google.com` [last visit: 15th May 2015]
13. `http://www.lec.com/power-translator-software.asp` [last visit: 15th May 2015]
14. `http://www.apertium.org` [last visit: 15th May 2015]

Language pair	Apertium	Google Translate	Power Translator
German→English (`de`→`en`)		✓	✓
English→German (`en`→`de`)		✓	✓
English→Spanish (`en`→`es`)	✓	✓	✓
Spanish→English (`es`→`en`)	✓	✓	✓
English→Finnish (`en`→`fi`)		✓	
Finnish→English (`fi`→`en`)		✓	
English→French (`en`→`fr`)		✓	✓
French→English (`fr`→`en`)		✓	✓
Spanish→French (`es`→`fr`)	✓	✓	✓
French→Spanish (`fr`→`es`)	✓	✓	✓

Table 1: MT systems available for each translation direction (→) used in the experiments.

CAT scenario, a prototype[15] plug-in implementing the training-free approach was built for the free/open-source CAT system OmegaT[16] and, after some experiments using the on-line MT systems Apertium and Google Translate, we can confirm that recommendations are obtained almost instantaneously.

6.1 Evaluation

The FMS-only baseline, the statistical alignment-based approach proposed in Section 3, and the MT-based approaches proposed in Sections 4, and 5 were tested by using a test set (TS) of parallel segments $\{(S'_l, T'_l)\}_{l=1}^{N}$, and a TM in the same domain. For each SL segment S'_l in TS, the set of matching TUs $\{(S_i, T_i)\}_{i=1}^{N'}$ in the TM with a FMS above threshold Θ is obtained. Please recall that the FMS measures the similarity between the translation proposals and the segments to be translated. The FMS threshold Θ is usually set to values above 60% (Bowker, 2002, p. 100) and in our experiments we therefore used several values of Θ of between 60% and 90%.[17] Once the set of matching TUs has been obtained, the recommendations for every word t_j in every target-language segment T_i are obtained and evaluated by using T'_l, the translation of S'_l, as a *gold standard*. The words in T'_l and T_i are matched by using the Levenshtein edit distance (Levenshtein, 1966), which allows us to check whether or not a given word t_j, the j-th word of T_i, is actually kept in the final translation. It is thus possible to determine whether a recommendation for t_j is successful both if:

- t_j is recommended to be changed in T_i and it does not match any word in T'_l, or
- t_j is recommended to be kept in T_i and it does match a word in T'_l.

15. `http://www.dlsi.ua.es/~mespla/edithints.html` [last visit: 15th May 2015]
16. `http://www.omegat.org` [last visit: 15th May 2015]
17. For those MT-based approaches that require training, different models were trained for every value of Θ included in the experiments.

Once all the pairs $(S'_l, T'_l) \in$ TS have been used to obtain their corresponding sets of matching TUs $\{(S_i, T_i)\}_{i=1}^{N'}$, and all the recommendations have been obtained and checked, several metrics are used for evaluation. Accuracy (A) is computed as the fraction of successful recommendations out of the total number of words for which a recommendation was made. It is worth noting that most of the methods proposed do not provide recommendations for all the words; another interesting metric is therefore the fraction of words not covered (NC) by the system, that is, the fraction of words for which no recommendation is made. The combination of these two metrics helps us to understand how each method perform on each test set. In addition to the accuracy and the fraction of words not covered, we also compute the precision and recall as regards keeping recommendations (P_K and R_K, respectively) and changing recommendations (P_C and R_C, respectively). The latter metrics are useful since they provide specific information about the successful keeping recommendations and change recommendations, separately, while A and NC provide information about the general performance of the recommender. The code used to perform these experiments is freely available under license GNU General Public License v3.0 (Free Software Fundation, 2007) and can be downloaded from `http://transducens.dlsi.ua.es/~mespla/resources/wkr/`.

6.2 Corpora

The corpus used in our experiments is the DGT-TM (Steinberger et al., 2012). This translation memory is a collection of documents from the *Official Journal of the European Union*[18] which is aligned at the segment level for several languages (multilingual TUs). Segment alignment in DGT-TM is expected to have a high level of quality, since part of these alignments were manually checked, or were actually generated during computer-aided translation by professional translators.

The TUs in DGT-TM contain segments in many official languages of the European Union and are labelled with domain codes[19] which were used to create three domain-specific TU collections. This was done by using the following domain codes: *elimination of barriers to trade* (code 02.40.10.40), *safety at work* (code 05.20.20.10), and *general information of public contracts* (code 06.30.10.00). Only those TUs containing the corresponding segments for all the five languages used in our experiments were included in these TU collections.

Each collection of TUs was used to build a bilingual TM and a test set for each language pair by randomly selecting pairs of segments without repetition.[20] In addition to the pre-processing already performed by the creators of DGT-TM (European Commission Joint Research Center, 2007) the segments included in the TMs and the test set used in our experiments were tokenised and lowercased. The TMs consist of $6,000$ TUs each, and simulate the TM that a translator may use when translating with a CAT tool. The test set consist of $1,500$ TUs whose source language side simulates the segments to be translated by using the TMs (the translator's job), while their target language side may be considered as a reference translation for each segment to be translated.

18. `http://eur-lex.europa.eu` [last visit: 15th May 2015]
19. `http://old.eur-lex.europa.eu/RECH_repertoire.do` [last visit: 15th May 2015]
20. The TMs and the test set obtained in this way can be downloaded from `http://transducens.dlsi.ua.es/~mespla/resources/mtacat/` [last visit: 15th May 2015]

	02.40.10.40	**05.20.20.10**	**06.30.10.00**
02.40.10.40	**0.99 (8°)**	0.24 (76°)	0.20 (78°)
05.20.20.10	0.26 (75°)	**0.98 (11°)**	0.23 (76°)
06.30.10.00	0.24 (76°)	0.23 (76°)	**0.97 (14°)**

Table 2: Cosine similarity (and the corresponding angle) between the English side of the English–Spanish TMs belonging to the three domains in the experiments: *elimination of barriers to trade* (code 02.40.10.40), *safety at work* (code 05.20.20.10), and *general information of public contracts* (code 06.30.10.00).

The domains chosen for the experiments have little overlap in vocabulary, as evidenced by the cosine similarity measure shown in Table 2.[21] This technique maps a text onto a *vocabulary vector*, in which each word is a dimension and the number of occurrences of this word in the text is the value for that dimension. These vocabulary vectors can be used to compare two texts by computing the cosine of the angle between them. The cosine similarity was computed using the English side of the three English–Spanish TMs and by splitting the 6,000 segments into two halves. The table shows the cosine similarity between the first half of each domain (rows) and the second half (columns).

As will be noted, the cosines between the vocabulary vectors from the same domain are very close to 1, with angles of between 8° and 14°. However, the cosines between the vocabulary vectors from different domains are much smaller, with angles of between 75° and 79°. We can therefore conclude that there are considerable differences between the TMs used in our experiments.

As regards the number of TUs matched when simulating the translation of Spanish segments into English in the test set, Table 3 reports, for fuzzy-match scores in four different ranges, the average number of TUs matched per segment to be translated and the total number of words for which to provide a recommendation. These data provide an idea of the repetitiveness of the corpora used to carry out the experiments. As can be seen, the corpus from domain 02.40.10.40 is more repetitive than the other two. It is worth noting that domains 05.20.20.10 and 06.30.10.00 have notable differences for low values of the FMS threshold Θ, while they do not differ that much for higher values.

6.3 Fuzzy-Match Score Function

For our experiments, as in many TM-based CAT systems, we have chosen a fuzzy-match score function based on the word-based Levenshtein edit distance (Levenshtein, 1966):

$$\mathrm{FMS}(S', S) = 1 - \frac{\mathrm{D}(S', S)}{\max(|S'|, |S|)}$$

21. The cosine similarity was computed on the lowercased corpora, removing punctuation characters and the stopwords provided in the 4.7.2 version of Lucene: `http://lucene.apache.org/core/` [last visit: 15th May 2015].

$\Theta(\%)$	**domain**	**no filtering**	
		TU_{avg}	N_{words}
≥ 60	02.40.10.40	3.71	95,881
	05.20.20.10	0.62	9,718
	06.30.10.00	5.68	34,339
≥ 70	02.40.10.40	2.36	65,865
	05.20.20.10	0.37	6,883
	06.30.10.00	0.99	10,327
≥ 80	02.40.10.40	1.58	46,519
	05.20.20.10	0.14	3,015
	06.30.10.00	0.45	4,726
≥ 90	02.40.10.40	0.70	26,625
	05.20.20.10	0.05	1,599
	06.30.10.00	0.03	1,268

Table 3: Average number of matching TUs (TU_{avg}) per segment and total number of target words (N_{words}) for which a recommendation has to be provided when translating Spanish into English for the three different domains. The results were obtained for different values of the FMS threshold (Θ).

where $|x|$ is the length (in words) of string x and $\text{D}(x, y)$ refers to the word-based Levenshtein edit distance between x and y.[22]

6.4 Binary Classifier

Esplà-Gomis et al. (2011) used a simple perceptron classifier which defined, for the translation of a source segment S', the probability of keeping the j-th word in T, the target-language segment of the TU (S, T) as:

$$p_k(j, S', S, T) = \frac{1}{1 + e^{-g(j,S',S,T)}} \tag{4}$$

with

$$g(j, S', S, T) = \lambda_0 + \sum_{k=1}^{N_F} \lambda_k f_k(j, S', S, T). \tag{5}$$

This perceptron uses a sigmoid function that incorporates the linear combination of the different features f_k and the corresponding weights λ_k learned by the classifier.

22. Many TM-based CAT tools implement variations of this FMS to rank the translation proposals as regards the edition effort required (for instance, by disregarding punctuation signs or numbers in S and S', or using stemmed versions of S and S'). In our experiments we continue to use the original FMS, since ranking is not important in our experiments. This is owing to the fact that all the proposals above the threshold are evaluated, and not only that with the highest score.

In this work, a more complex *multilayer perceptron* (Duda, Hart, & Stork, 2000, Section 6) was used, namely, that implemented in Weka 3.7 (Hall et al., 2009). Multilayer perceptrons (also known as feedforward neural networks) have a complex structure which incorporates one or more *hidden layers*, consisting of a collection H of perceptrons, placed between the input of the classifier (the features) and the output perceptron. This hidden layer makes multilayer perceptrons suitable for non-linear classification problems (Duda et al., 2000, Section 6). In fact, Hornik, Stinchcombe, and White (1989) proved that neural networks with a single hidden layer containing a finite number of neurons are universal approximators and may therefore be able to perform better than a simple perceptron for complex problems. In this case, the output perceptron that provides the classification takes the output h_l of each of the perceptrons in H as its input. Eq. (5) therefore needs to be updated as follows:

$$g(j, S', S, T) = \lambda_0 + \sum_{l=1}^{|H|} \lambda_l h_l(j, S', S, T). \tag{6}$$

Each perceptron h_l in H works similarly to the perceptron described in eq. (4):

$$h_l(j, S', S, T) = \frac{1}{1 + e^{-g_l(j,S',S,T)})}$$

with

$$g_l(j, S', S, T) = \lambda_{l0} + \sum_{k=1}^{N_F} \lambda_{lk} f_k(j, S', S, T).$$

As can be seen, besides the collection of weights $\boldsymbol{\lambda}$ for the main perceptron, a different collection of weights $\boldsymbol{\lambda}'_l$ is needed for each perceptron h_l in the hidden layer H. These weights are obtained by using the *backpropagation algorithm* (Duda et al., 2000, Section 6.3) for training, which updates them using gradient descent on the error function. In our case, we have used a batch training strategy, which iteratively updates the weights in order to minimise an error function. The training process stops when the error obtained in an iteration is worse than that obtained in the previous 10 iterations.[23]

A validation set with 10% of the training examples was used during training, and the weights were therefore iteratively updated on the basis of the error computed in the other 90%, but the decision to stop the training (usually referred as the convergence condition) was based on this validation set. This is a usual practice whose objective is to minimise the risk of overfitting.

Hyperparameter optimisation was carried out using a grid search (Bergstra, Bardenet, Bengio, & Kégl, 2011) strategy based on the accuracy obtained for the English–Spanish TM from the 02.40.10.40 domain. A 10-fold cross-validation was performed on this training corpus in order to choose the following hyperparameters:

23. It is usual to set a number of additional iterations after the error stops improving, in case the function is in a local minimum, and the error starts decreasing again after a few more iterations. If the error continues to be worsen after these 10 iterations, the weights used are those obtained after the iteration with the lowest error.

- *Number of nodes in the hidden layer*: Weka (Hall et al., 2009) makes it possible to choose from among a collection of predefined network designs; that which best performed for our training corpus was that with the same number of nodes in the hidden layer as the number of features.
- *Learning rate*: this parameter allows the dimension of the weight updates to be regulated by applying a factor to the error function after each iteration; the value that best performed for our experiment was 0.4.
- *Momentum*: when updating the weights at the end of a training iteration, momentum modifies the new value, signifying that it not only depends on the current gradient direction, but also on the previous weight value. The objective of this technique is to smooth the training process for faster convergence. In the case of our experiments, it was set to 0.1.

6.5 Reference Results

As mentioned previously, the performance of the two MT-based approaches proposed in this work is compared to that of two different approaches: a *naïve FMS-only baseline*, which uses the classifier described in Section 6.4 and employs only the FMS between S' and S as a feature, and the approach reviewed in Section 3, which uses statistical word alignment to relate the words in the two segments of a TU (S, T). The naïve FMS-only baseline was trained on the datasets described in Section 6.2 for different values of the FMS threshold Θ. It is worth mentioning that the resulting models classify all the target words as having to be kept. This is a consequence of the fact that, for any value of the FMS in the training set, there are more words to be kept than to be changed.

The alignments used by the alignment-based approach were obtained by means of the free/open-source MGIZA++ toolkit (Gao & Vogel, 2008), an implementation of the GIZA++ toolkit (Och & Ney, 2003) which eases the task of training alignment models on a parallel corpus and then aligning a different one using the models learned. The word-based alignment models (Brown et al., 1993; Vogel, Ney, & Tillmann, 1996) were separately trained on the TMs defined in Section 6.2 and on the JRC-Acquis 3.0 (Steinberger et al., 2006) corpus (a large multilingual parallel corpus which includes, among others, texts from these TMs, given that it is built from the same texts as the DGT-TM).[24] The alignments we have used are the result of running MGIZA++ in both translation directions (source-to-target and target-to-source) and then symmetrising both sets of alignments by means of the usual *grow-diag-final-and* (Koehn et al., 2005) heuristic. This symmetrisation technique was found to be that which provided the best compromise between coverage and accuracy for word-keeping recommendation (Esplà et al., 2011).[25]

Table 4 shows the accuracy obtained by the naïve FMS-only baseline. The fraction of words not covered, that is, the words for which no recommendation is provided, is not included in this table since this baseline provides a recommendation for every word in the

24. `https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis` [last visit: 15th May 2015]
25. Symmetrisation is necessary because MGIZA++ 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. The use of symmetrisation allows alignments to be combined in both directions in order to obtain M to N alignments.

$\Theta(\%)$	$A(\%)$
≥ 60	82.69±.24
≥ 70	88.37±.25
≥ 80	91.65±.25
≥ 90	93.98±.29

Table 4: Accuracy $A(\%)$ obtained with the naïve FMS-only baseline when translating en→es in domain 02.40.10.40. Accuracy was obtained for different FMS thresholds Θ. The other language pairs and domains behave in the same way.

test set. This is due to the fact that the naïve FMS-only baseline does not depend on the coverage of a source of information.

In general, we can see that the accuracy obtained with the naïve FMS-only baseline is quite high. This is, in fact, a hard-to-beat naïve baseline, although these results are reasonable, since the relatively high values of the FMS threshold Θ imply that a high number of words should be kept unedited in the translation proposals.

With regard to the alignment-based approach, several options were evaluated in order to choose its configuration. On the one hand, we tried the two decision criteria described in Section 3 (*unanimity* and *majority*). On the other hand, we tried two alignment models: one trained on the same translation memory used for the experiments (as had occurred with the trained MT-based recommender), and another trained on the JRC Acquis parallel corpus. The objective of comparing both models was to confirm which corpus was the most adequate as regards training our alignment-based recommender: one which was reduced and domain focused, or one which was bigger and more generic, although still containing text in the same domain. The results are presented in Table 5, in which the accuracy and percentage of words not covered are measured for the four combinations of decision criteria and training corpora. As already mentioned in Section 3, the unanimity criterion is more focused on accuracy, while the majority criterion is more focused on coverage. In order to confirm which method was better, a statistical significance test was performed on the results obtained by using an approximate randomisation test.[26] The free/open-source tool SIGF V.2 (Padó, 2006) was used for the statistical significance testing of the results described throughout this section. The test confirmed that in both cases the alignment models trained on the TM used for testing outperform those trained on the JRC Acquis corpus. The approach trained on the TM used for testing will be used in all the experiments shown in the following section, while the decision criterion used will be that of unanimity, the reason being that we consider accuracy to be more relevant than coverage for word-keeping recommendation, since as mentioned above, we believe that it is better not to make a recommendation than to make a wrong one.

26. Approximate randomisation compares the difference between the accuracy/coverage of two classifiers in the same test set. This method randomly interchanges the predictions of both classifiers in every instance of the test set. The difference between the accuracy/coverage in both randomised datasets is then compared to the original set. This process is iteratively repeated to confirm whether the results as regards randomised predictions are consistently worse than the original results.

Θ(%)	method	training on 02.40.10.40		training on JRC Acquis	
		A (%)	NC (%)	A (%)	NC (%)
≥ 60	unanimity	**93.90±.16**	6.09±.15	93.14±.17	6.29±.15
	majority	92.96±.17	**4.39±.13**	93.05±.17	5.39±.14
≥ 70	unanimity	**94.32±.18**	5.90±.18	93.67±.19	6.15±.18
	majority	93.47±.19	**4.45±.16**	93.59±.19	5.42±.17
≥ 80	unanimity	**95.10±.20**	5.37±.21	94.56±.21	5.87±.21
	majority	94.49±.21	**4.31±.19**	94.52±.21	5.33±.20
≥ 90	unanimity	**95.34±.26**	4.93±.26	94.97±.27	5.48±.27
	majority	95.10±.27	**4.35±.25**	94.95±.27	5.04±.26

Table 5: Accuracy (A) and fraction of words not covered (NC) obtained with the alignment-based approach described in Section 3 for different FMS thresholds Θ, when translating Spanish into English in domain 02.40.10.40. The results show the accuracy obtained when using a model trained on the TM belonging to the 02.40.10.40 domain and on the JRC-Acquis corpus, both using the unanimity and the majority decision criteria. This behaviour is also observed for the remaining TMs used in the experiments. Statistically significant differences in the accuracy of each approach for the different values of Θ with $p \leq 0.05$ are highlighted in bold type, as also occurs for the fraction of words not covered.

7. Results and Discussion

In this section we present the results obtained by the two approaches proposed in this paper and compare their performance with the naïve FMS-only baseline and the alignment-based approach of Esplà et al. (2011). The large amount of variables to be taken into consideration (feature sets, language pairs, domains, MT systems, and sub-segment length) forced us to select the experiments to be performed. Some parameters were therefore chosen on the basis of the results obtained for the translation from Spanish into English, which is the language pair used by Esplà et al. (2011) and Esplà-Gomis et al. (2011). The domain chosen for these preliminary experiments is *elimination of barriers to trade* (02.40.10.40), which has higher matching rates (see Table 3) and is therefore that from which more data can be obtained.

7.1 Parameter Selection

We first attempted to determine the optimal sub-segment maximum length L for the experiments with the training-free recommender and with the trained recommender. Table 6 shows the fraction of words not covered depending on the value of L for both recommenders together. The fraction of words not covered is between 16% and 19% when using sub-segments of only one word, and the percentage diminishes as more context is provided for translations. As can be seen, the fraction of words not covered starts to stabilise with $L = 4$, since the difference between this and $L = 5$ is only about 0.25%.

Table 7 shows the impact of the value of L on the accuracy obtained by the training-free recommender and by the trained recommender when using the different sets of features

$\Theta(\%)$	**Fraction of words without recommendation (%)**				
	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$
≥ 60	16.42±.24	10.22±.19	7.24±.16	5.13±.14	4.90±.14
≥ 70	16.74±.28	10.66±.24	7.34±.20	5.18±.17	4.94±.17
≥ 80	17.37±.34	11.25±.29	7.65±.24	5.53±.21	5.29±.20
≥ 90	18.18±.46	11.80±.39	8.05±.33	5.86±.28	5.59±.28

Table 6: Percentage of words not covered by both MT-based approaches for the en→es language pair in domain 02.40.10.40 using a combination of all MT systems available. The fraction of words not covered was obtained for different FMS thresholds Θ when using different values of the maximum sub-segment length L.

$\Theta(\%)$	**Method**	**Accuracy (%) in classification**				
		$L=1$	$L=2$	$L=3$	$L=4$	$L=5$
≥ 60	**MM-U**	93.51±.17	93.40±.17	93.58±.16	93.57±.16	93.77±.16
	PM-C	93.62±.17	94.07±.16	94.31±.15	94.18±.15	94.37±.15
	PM-C+C	93.59±.17	94.36±.15	94.57±.15	**95.14±.14**	95.41±.14
	training-free	93.63±.17	93.78±.16	93.79±.16	93.27±.16	92.90±.17
≥ 70	**MM-U**	94.79±.19	94.72±.18	94.77±.18	94.70±.18	94.82±.17
	PM-C	94.75±.19	94.89±.18	95.12±.17	94.94±.17	95.05±.17
	PM-C+C	94.81±.19	95.16±.17	95.33±.17	**95.63±.16**	95.92±.16
	training-free	94.76±.19	94.77±.18	94.77±.18	94.14±.18	93.78±.19
≥ 80	**MM-U**	96.09±.19	96.14±.19	95.92±.19	96.00±.18	96.02±.18
	PM-C	96.09±.19	96.14±.19	96.11±.18	96.01±.18	95.98±.18
	PM-C+C	96.11±.19	96.24±.18	96.34±.18	**96.39±.17**	96.58±.17
	training-free	96.05±.20	95.97±.19	95.88±.19	95.29±.20	94.98±.20
≥ 90	**MM-U**	96.84±.23	96.85±.22	96.80±.22	96.74±.22	96.75±.22
	PM-C	96.84±.23	96.82±.23	96.87±.22	96.85±.22	96.87±.22
	PM-C+C	96.84±.23	96.95±.22	96.95±.22	**96.90±.21**	97.00±.21
	training-free	96.80±.23	96.72±.23	96.70±.22	96.61±.22	96.42±.23

Table 7: Accuracy obtained by the trained MT-based recommender when using the different feature combinations described in Section 4 and by the training-free MT-based recommender for the en→es language pair. Accuracy was obtained for different FMS thresholds Θ when using different values of the maximum sub-segment length L. Statistically significant accuracy results for $L = 4$ (the value of L that will be used for the remaining experiments) with $p \leq 0.05$ are highlighted in bold type.

described in Section 4. As can be seen, the accuracy of the training-free system drops slightly as longer sub-segments are introduced. This is reasonable since the longer the sub-segments used, the higher the number of words for which a recommendation is made (see Table 6). Words which are covered only by very long sub-segments are more difficult to classify, since these sub-segments contain evidence regarding more words and are therefore less precise. It is interesting to observe that, in the case of most of the feature sets, the trained recommender does not behave in this manner, since it is able to learn how reliable the longer sub-segments are. In the case of the feature set MM-U, the accuracy is almost constant for all the values of L, which means that using longer sub-segments does not have an impact on the accuracy in this case. In any event, it is worth noting that the results obtained using the training-free recommender are quite accurate, which confirms that the sub-segment pairs discovered using MT are a good source of information for word-keeping recommendation. Moreover, these results indicate that long sub-segments are less informative than short sub-segments. In general, only small improvements in accuracy and coverage occur for values of L that are higher than 4. The remaining experiments in this section will therefore be performed with $L = 4$.

The results in Table 7, namely those in the column with $L = 4$, were also used to determine which is the best feature combination for the trained MT-based recommender. At first glance, the set of features based on matching words, namely PM-C (see Section 4.2), and PM-C+C (see Section 4.3), are those which perform best. As commented on in Section 4.2, MM-U features consider partial matching sub-segments as negative evidence, while PM-C and PM-C+C also attempt to extract the positive evidence from these sub-segments, thus using this bilingual information more efficiently. However, the results that they obtain are very close, particularly in the case of high values of Θ. A statistical significance test confirmed that PM-C+C is superior to all the other feature combinations for any value of Θ with $p \leq 0.05$. The feature set PM-C+C will therefore be used for the trained classifier approach in the remaining experiments in this section.

The accuracy obtained using all the approaches presented in both this and the previous section is lower than expected, particularly when considering the results obtained by Esplà-Gomis et al. (2011) and Esplà et al. (2011) in which both the trained MT-based approach and the alignment-based approach obtained an average accuracy that was about 5% higher than that obtained in our experiments. Our intuition leads us to believe that this drop in accuracy may be due to the fact that the data sets used in previous works might have been cleaner than those used here. To confirm this, an additional set of experiments was carried out using some additional cleaning criteria to ensure the quality of the datasets used for evaluation. The results of this study are presented in Appendix B.

7.2 General Results

The parameters chosen (maximum sub-segment length $L = 4$ for the MT-based approaches, PM-C+C feature set for the trained MT-based approach, and unanimity criterion and models trained on the TM to be used in the experiments for the alignment-based approach) were used to perform several experiments in order to check the performance of our system. Tables 8 and 9 show the results obtained by the trained MT-based recommender when translating Spanish segments into English. In Table 8, the different MT systems available

Θ(%)	Apertium		Google Translate		Power Translator	
	A (%)	NC (%)	A (%)	NC (%)	A (%)	NC (%)
≥ 60	94.61±.17	27.89±.28	**95.08±.14**	**6.22±.15**	94.52±.17	28.99±.29
≥ 70	95.40±.19	28.32±.34	95.54±.16	**6.38±.19**	95.47±.19	29.62±.35
≥ 80	96.49±.20	28.39±.41	96.34±.18	**6.52±.22**	96.46±.20	29.59±.42
≥ 90	**97.25±.23**	28.49±.54	96.99±.21	**6.31±.29**	97.01±.24	28.50±.54

Table 8: Accuracy (A) and fraction of words not covered (NC) obtained when translating es→en in domain 02.40.10.40 with the trained MT-based approach. The results were obtained with the separate use of the MT systems available for the language pair: Apertium, Google Translate, and Power Translator. For every value of Θ, those results that supersede the rest by a statistically significant margin of $p \leq 0.05$ are highlighted in bold type, both for accuracy and for the fraction of words not covered.

were used separately to obtain the recommendations in the 02.40.10.40 domain. The results confirm that, while accuracy remains stable,[27] coverage strongly depends on the MT system. This may be interpreted as follows: MT-based approaches are robust to bilingual sources of information with low coverage. The experiments confirmed that the best coverage is obtained with Google Translate, whereas Apertium and Power Translator produce similar results. However, Apertium and Power Translator produce higher precision for "change" recommendations, while all three MT systems perform similarly as regards the precision for "keep" recommendations.

Table 9 compares the performance of the alignment-based approach and the trained MT-based recommender when all three MT systems available are used at the same time for the language pair es→en. The table shows the results obtained separately for the alignment-based approach and the trained MT-based approach as regards the three domains: 02.40.10.40, 05.20.20.10 and 06.30.10.00. The results are quite similar for both approaches. The MT-based approach slightly outperforms the alignment-based approach in accuracy, while the results of the alignment-based approach are better for coverage, particularly in the case of the 06.30.10.00 domain. In any event, this leads us to believe that both approaches can obtain comparable results across domains.

Tables 10 and 11 present the results as regards the accuracy and fraction of words not covered, respectively, obtained with both the trained MT-based approach and the alignment-based approach for all language pairs. In the case of the trained MT-based approach, all the MT systems available were used (Table 1 lists the MT systems available for each language pair).

The results confirm the hypothesis that the alignment-based approach generally obtains better results as regards coverage than the MT-based approach for most language pairs, which is reasonable, given that the alignment models have been trained on the same TM to be used for translation. Accuracy is yet again the strongest point of the trained MT-

27. Although some differences in accuracy can be observed, it is only possible to state which MT system is better with a statistical significance of $p \leq 0.05$ in the case of Θ=60% and Θ=90%.

Θ(%)	method	02.40.10.40		05.20.20.10		06.30.10.00	
		A (%)	NC (%)	A (%)	NC (%)	A (%)	NC (%)
≥ 60	trained	**95.14±.1**	**5.13±.1**	**95.62±.4**	7.20±.5	**94.87±.3**	**11.33±.3**
	alignment	93.90±.2	6.09±.2	92.97±.5	**6.22±.5**	91.38±.3	18.31±.4
≥ 70	trained	**95.63±.2**	**5.18±.2**	**97.02±.4**	6.44±.6	94.93±.5	10.55±.6
	alignment	94.32±.2	5.90±.2	94.16±.6	5.96±.6	94.57±.5	**7.78±.5**
≥ 80	trained	**96.39±.2**	5.53±.2	**95.78±.8**	10.42±1.1	95.00±.7	12.84±1
	alignment	95.10±.2	5.37±.2	94.56±.8	**5.51±.8**	95.20±.6	**4.80±.6**
≥ 90	trained	**96.90±.2**	5.86±.3	95.36±1.1	11.13±1.5	97.65±.9	9.31±1.6
	alignment	95.34±.3	**4.93±.3**	94.26±1.2	**5.19±1.1**	96.47±1.1	**6.15±1.3**

Table 9: Accuracy (A) and fraction of words not covered (NC) obtained with the trained MT-based approach and the alignment-based approach when translating es→en in the three different domains and using all the MT systems. For each corpus and for each value of Θ, a statistical significance test was performed for both approaches as regards both the accuracy and the fraction of words not covered. Those results that show an improvement which is statistically significant with $p \leq 0.05$ are highlighted in bold type.

	Θ ≥ 60(%)		Θ ≥ 70(%)		Θ ≥ 80(%)		Θ ≥ 90(%)	
lang. pair	trained	alignment	trained	alignment	trained	alignment	trained	alignment
es→en	**95.1±.1**	93.9±.2	**95.5±.2**	94.3±.2	**96.3±.2**	95.1±.2	**97.0±.2**	95.3±.3
en→es	**90.0±.2**	88.7±.2	**91.5±.2**	89.8±.2	**92.3±.2**	90.4±.2	**93.0±.2**	91.8±.3
de→en	**94.2±.2**	92.6±.2	**95.8±.2**	93.8±.2	**96.6±.2**	94.7±.2	**97.2±.2**	95.5±.3
en→de	**88.9±.2**	87.9±.2	**91.1±.2**	89.7±.2	**93.0±.2**	91.5±.2	**92.8±.3**	91.5±.3
fr→en	**95.2±.1**	93.4±.2	**96.3±.2**	95.1±.2	**97.0±.2**	96.0±.2	**97.6±.2**	96.4±.2
en→fr	**89.7±.2**	89.4±.2	**91.9±.2**	91.3±.2	**93.3±.2**	92.1±.2	**95.7±.2**	94.3±.3
fi→en	**93.2±.2**	91.9±.2	**94.6±.2**	93.1±.2	**94.7±.2**	93.3±.3	**94.8±.3**	92.8±.4
en→fi	**89.1±.2**	87.1±.2	**90.2±.3**	88.5±.3	**90.3±.3**	88.4±.3	**90.6±.4**	88.9±.4
es→fr	**91.2±.2**	89.4±.2	**93.2±.2**	90.7±.2	**94.8±.2**	92.8±.2	**96.0±.2**	94.3±.2
fr→es	**89.4±.2**	88.6±.2	**92.1±.2**	91.1±.2	**93.5±.2**	92.2±.2	**93.6±.3**	92.2±.3

Table 10: Accuracy (A) obtained with both the trained MT-based approach and the alignment-based approach when translating between all the language pairs in domain 02.40.10.40. The results were obtained for several values of the FMS threshold Θ and using all available MT systems for each language pair. For each language pair and for each value of Θ, a statistical significance test was performed between the accuracy obtained by both approaches. Those results that show an improvement which is statistically significant with $p \leq 0.05$ are highlighted in bold type.

lang. pair	Θ ≥ 60(%) trained	Θ ≥ 60(%) alignment	Θ ≥ 70(%) trained	Θ ≥ 70(%) alignment	Θ ≥ 80(%) trained	Θ ≥ 80(%) alignment	Θ ≥ 90(%) trained	Θ ≥ 90(%) alignment
es→en	6.2±.1	6.1±.1	6.4±.2	**5.9±.2**	6.5±.2	**5.4±.2**	6.3±.3	**4.9±.3**
en→es	**7.3±.1**	7.7±.1	8.3±.2	8.8±.2	**8.6±.2**	9.4±.2	**9.5±.3**	10.7±.3
de→en	10.5±.2	**5.8±.1**	10.7±.2	**6.2±.2**	10.9±.3	**6.0±.2**	12.4±.4	**6.0±.3**
en→de	11.6±.2	**5.6±.1**	12.4±.2	**6.2±.2**	13.0±.3	**6.5±.2**	14.5±.4	**6.6±.3**
fr→en	7.7±.2	**4.3±.1**	7.7±.2	**4.1±.2**	8.1±.3	**4.2±.2**	7.6±.3	**3.5±.2**
en→fr	7.9±.1	**5.9±.1**	8.8±.2	**6.0±.2**	9.6±.2	**6.0±.2**	9.4±.3	**7.8±.3**
fi→en	11.2±.2	**9.9±.2**	11.3±.3	**10.2±.2**	11.5±.3	**10.6±.3**	11.6±.4	**10.9±.4**
en→fi	11.6±.2	**7.2±.2**	11.0±.3	**7.0±.2**	12.0±.3	**7.2±.2**	12.6±.4	**8.0±.3**
es→fr	17.8±.2	**4.9±.1**	18.4±.2	**4.9±.1**	18.6±.3	**5.0±.2**	21.2±.4	**5.6±.2**
fr→es	19.5±.2	**10.3±.1**	19.4±.3	**9.5±.2**	20.4±.3	**10.5±.2**	17.6±.4	**5.9±.2**

Table 11: Fraction of words not covered (NC) obtained with both the trained MT-based approach and the alignment-based approach when translating between all the language pairs in domain 02.40.10.40. The results were obtained for several values of the FMS threshold Θ and using all available MT systems for each language pair. For each language pair and for each value of Θ, a statistical significance test was performed between the fraction of words not covered obtained by both approaches. Those results that show an improvement which is statistically significant with $p \leq 0.05$ are highlighted in bold type.

based recommender. Another interesting detail is that the experiments for language pairs with English as the target language obtain better accuracy than the experiments for the inverse language pairs. This is due to the fact that in the DGT-TM, it is usual to find *free translations* in languages other than English, which is the original language of most of the documents in this TM.[28] It is particularly frequent to find additional information about technical English words in the other languages. For example, when *software* is translated into Spanish, the translated segments include the text *"equipo lógico" ("software").* The translation includes both the correct translation and the original word, in order to keep the meaning of the original English segment. This is an important issue since, when these free translations are used as a reference to evaluate the accuracy of the approaches presented in this work, they lead to lower accuracy. This problem is analysed, and partially bypassed, in the additional experiments presented in Appendix B. As can be seen, the trained MT-based approaches provide, in general, better accuracy. However, for most of the language pairs, the coverage obtained with the alignment-based approach is much better. Nevertheless the results are still reasonably similar and can be considered comparable.

Table 12 shows the results obtained with the different approaches:[29] the naïve FMS-only baseline, the alignment-based approach, and both the trained and the training-free MT-based approaches, for the es→en language pair in the 02.40.10.40 domain and using all the MT systems available simultaneously. This table provides more detail: in addition to the accuracy and percentage of words not covered, it also includes the precision and recall for both keeping recommendations and changing recommendations. This table allows to better understand the differences between the approaches before starting a more complex comparison in different scenarios. Throughout this section we provide information regarding precision and recall where it is significant for all the tables presented.

Leaving aside the naïve FMS-only baseline, it can be observed that the accuracy is similar for all the approaches, and that the trained MT-based recommender obtains slightly better results. As already mentioned, the amount of words not covered is very similar for the MT-based approaches and the alignment-based approach. As regards precision, the trained MT-based approach seems to outperform the others, although all the approaches obtain comparable scores. These results are coherent with those obtained for the rest of language pairs: in general recall and precision in "keep" recommendations are similar for both approaches, while the MT-based approach seems to be more precise in the case of "change" recommendations, where the difference is much higher, specially for higher values of the FMS threshold Θ. These conclusions are extensible to the data shown in Table 13.

The results shown in Table 12 were extended by repeating the experiment and computing recommendations only for content words (i.e. ignoring stopwords). This was done by using the list of stopwords provided in the 4.7.2 version of Lucene[30] for the language utilised in our experiments. The results of this experiments can be found in Table 13. As can be seen, the results do not change much, thus confirming that all the approaches perform equally well with content and stopwords.

28. According to Steinberger et al. (2012), English is the source language in 72% of the documents.
29. The naïve FMS-only baseline did not recommend that any word be changed, as explained in Section 6.5, signifying that P_c cannot be computed for this approach and R_c is always 0. Similarly, all the words in this approach are covered, and therefore R_k is always 1.
30. `http://lucene.apache.org/core/` [last visit: 15th May 2015]

Θ(%)	Method	A (%)	NC (%)	P_{k} (%)	R_{k} (%)	P_{c} (%)	R_{c} (%)
≥ 60	**FMS-only**	82.7±.2	100%±0	82.7±.2	100%±0	—	0%±0
	alignment	93.9±.2	6.1±.2	96.3±.1	96.5±.1	80.8±.3	80.2±.3
	trained	**95.1±.1**	**5.1±.1**	96.5±.1	**97.8±.1**	**87.6±.2**	**81.8±.3**
	training-free	93.3±.2	**5.1±.1**	95.2±.1	96.8±.1	82.2±.3	74.9±.3
≥ 70	**FMS-only**	88.4±.3	100%±0	88.4±.3	100%±0	—	0%±0
	alignment	94.3±.2	5.9±.2	96.6±.1	97.2±.1	73.0±.4	69.1±.4
	trained	**95.6±.2**	**5.2±.2**	**96.8±.1**	**98.4±.1**	**83.7±.3**	**71.8±.4**
	training-free	94.1±.2	**5.2±.2**	95.9±.2	97.7±.1	75.8±.3	63.6±.4
≥ 80	**FMS-only**	91.7±.3	100%±0	91.7±.3	100%±0	—	0%±0
	alignment	95.1±.2	5.4±.2	96.8±.2	98.0±.1	68.4±.4	57.2±.5
	trained	**96.4±.2**	5.5±.2	**97.2±.2**	**99.0±.1**	**80.8±.4**	**61.1±.5**
	training-free	95.3±.2	5.5±.2	96.5±.2	98.5±.1	71.0±.4	51.3±.5
≥ 90	**FMS-only**	94.0±.3	100%±0	94.0±.3	100%±0	—	0%±0
	alignment	95.3±.3	**4.9±.3**	96.5±.2	98.7±.1	57.1±.6	**32.4±.6**
	trained	**96.9±.2**	5.9±.3	**97.3±.2**	**99.6±.1**	**72.9±.6**	30.1±.6
	training-free	96.6±.2	5.9±.3	**97.4±.2**	99.2±.1	60.2±.6	**32.6±.6**

Table 12: Comparison of the results obtained using the trained MT-based approach, the training-free MT-based approach, the alignment-based approach and the naïve FMS-only baseline. The accuracy (A) and fraction of words not covered (NC) are reported, together with the precision (P) and recall (R) as regards both keeping recommendations and changing recommendations. The results were obtained for several values of the FMS threshold Θ when translating es→en in domain 02.40.10.40, using all the MT systems available. Statistically significant results with $p \leq 0.05$ are highlighted in bold type. For some values of Θ two values are highlighted in the same column; this means that there is not a statistically significant difference between these results, but both of them are significantly better than the other values.

Θ(%)	Method	A (%)	NC (%)	P_{k} (%)	R_{k} (%)	P_{c} (%)	R_{c} (%)
≥ 60	**FMS-only**	81.2±.3	100%±0	81.2±.3	100%±0	—	0%±0
	alignment	93.7±.2	6.2±.2	96.1±.1	96.4±.1	82.1±.3	80.8±.3
	trained	**95.1±.2**	**5.6±.2**	**96.5±.1**	**97.6±.1**	**88.1±.2**	**83.7±.3**
	training-free	93.3±.2	**5.6±.2**	95.3±.2	96.6±.1	83.0±.3	77.9±.3
≥ 70	**FMS-only**	87.3±.3	100%±0	87.3±.3	100%±0	—	0%±0
	alignment	94.0±.2	5.9±.2	96.3±.2	97.0±.1	74.0±.4	69.6±.4
	trained	**95.6±.2**	5.7±.2	**96.9±.2**	**98.2±.1**	**84.3±.3**	**74.9±.4**
	training-free	93.9±.2	5.7±.2	95.9±.2	97.3±.1	76.0±.4	67.1±.4
≥ 80	**FMS-only**	90.8±.3	100%±0	90.8±.3	100%±0	—	0%±0
	alignment	94.9±.2	**5.2±.2**	96.6±.2	97.9±.1	70.5±.5	58.7±.5
	trained	**96.4±.2**	6.1±.2	**97.3±.2**	**98.8±.1**	**81.6±.4**	**65.6±.5**
	training-free	95.1±.2	6.1±.2	96.5±.2	98.2±.1	71.5±.5	55.4±.5
≥ 90	**FMS-only**	93.4±.3	100%±0	93.4±.3	100%±0	—	0%±0
	alignment	94.9±.3	**4.5±.3**	96.1±.3	98.6±.2	58.9±.7	33.0±.7
	trained	**96.9±.2**	6.6±.3	**97.3±.2**	**99.5±.1**	**74.0±.6**	34.5±.7
	training-free	96.5±.3	6.6±.3	**97.4±.2**	99.0±.1	60.5±.7	**36.9±.7**

Table 13: Comparison of the results obtained using the trained MT-based approach, the training-free MT-based approach, the alignment-based approach and the naïve FMS-only baseline. The accuracy (A) and fraction of words not covered (NC) are reported, together with the precision (P) and recall (R) for both keeping recommendations and changing recommendations. Unlike Table 12, these metrics were computed on content words only. The results were obtained for several values of the FMS threshold Θ when translating es→en in domain 02.40.10.40, using all the MT systems available. Statistically significant results with $p \leq 0.05$ are highlighted in bold type. For some values of Θ two values are highlighted in the same column; this means that there is not a statistically significant difference between these results, but both of them are significantly better than the other values.

training language pair	Θ(%) ≥ 60	≥ 70	≥ 80	≥ 90
es→en	95.08±.14	95.54±.16	96.34±.18	96.99±.21
en→es	90.84±.19	91.80±.22	93.35±.23	94.94±.27
de→en	92.52±.17	93.46±.20	95.09±.20	96.41±.23
en→de	92.20±.18	93.22±.20	94.31±.22	94.88±.27
fr→en	92.65±.17	**94.16±.19**	**95.78±.19**	**96.67±.22**
en→fr	90.91±.19	92.23±.21	93.84±.23	95.24±.26
fi→en	92.38±.17	93.93±.19	95.30±.20	96.35±.23
en→fi	91.05±.19	93.36±.20	95.20±.20	96.47±.23
es→fr	92.36±.17	93.40±.20	94.56±.21	96.42±.23
fr→es	91.91±.18	92.24±.21	93.16±.24	95.49±.26
training-free	**93.36±.16**	**94.28±.18**	95.41±.20	**96.72±.22**

Table 14: Accuracy obtained by the trained MT-based recommender when translating Spanish into English in domain 02.40.10.40 by using recommendation models trained for other language pairs in the same domain. The first row, highlighted in grey, corresponds to the reference results obtained with the model trained on es→en. Only Google Translate was used for this experiment. Statistically significant results with $p \leq 0.05$ are highlighted in bold type. For some values of Θ two values are highlighted in the same column; this means that there is no statistically significant difference between these results, but both of them are significantly better than the other values.

The experiments carried out up to this point have confirmed that the three approaches proposed in this work perform similarly in different scenarios. While the word-alignment based approach provides the highest coverage and is, therefore, able to provide more recommendations, the MT-based approaches are more robust and obtain higher and more stable accuracy independently of the language pair or domain used. These results have led us to believe that no approach clearly stands out as being better, and that all of them may be useful in different scenarios, depending on the resources available and the translation conditions.

7.3 Experiments on Reusability Across Language Pairs

Table 14 presents the results obtained with the trained MT-based recommender when used in the same domain but with a different language pair. The experiments were performed in domain 02.40.10.40 when translating Spanish segments into English and re-using models trained on other language pairs. The results obtained with the model trained on this pair of languages are included in the table to give an idea of the upper-bound, and the corresponding row is filled in grey. For all the models used, both for training and testing, the only source of information used was Google Translate, since it is the only MT system which is available for all the language pairs used in our experiments.

The results show a clear decline as regards the results obtained when the recommendation model is learned for es→en and when it is learned from the other language pairs, particularly for low values of the FMS threshold Θ. In most cases, the accuracy obtained when training the recommender on language pairs that are different from those used for testing is worse than that obtained using the training-free approach. The only exception is the model trained on the fr→en pair, which is the most similar pair to es→en. The statistical significance test confirms that, for all the values of Θ, either this model or the training-free approach are the best ones. The difference between the accuracy obtained for these approaches for Θ=70% and for Θ=90% is not in fact statistically significant.

These results have led us to believe that the trained method is highly dependent on the language pair used for training, thus making it reasonable to conclude that it is better to use the training-free MT-based recommender than an MT-based recommender trained on a different language pair.

7.4 Experiments on Reusability Across Domains

Table 15 presents the results of the experiments concerning domain independence. The objective of these experiments is to verify how dependent the trained MT-based recommender is on the domain of the training corpus. In this case, we re-used the recommendation models trained in the three domains for the es→en translation to translate Spanish segments from domain 02.40.10.40 into English, and using all the MT systems available.

A drop in accuracy can be observed when re-using models trained on out-of-domain TMs rather than training on the TM to be used for translation. However, in this case the accuracy is closer to that obtained when the recommendation model is trained on the same TM used for testing (in-domain). With regard to the results obtained with the alignment-based approach, the difference in accuracy of all the MT-based approaches is higher and statistically significant with $p \leq 0.05$. That is to say, training-free MT-based approach and the models trained on domain 05.20.20.10 are those which perform best, with no statistically significant difference for most values of Θ. Similarly, the coverage of the alignment-based approach clearly drops when using out-of-domain models. This is due to the fact that, in the case of the alignment-based approach, those words which were not seen during training cannot be aligned, since no translation probabilities are learned for them. In contrast, in the case of the MT-based approaches the linguistic resources are not learned during training, but are rather obtained from the MT systems available: the training of the MT-based recommender instead focuses on the relevance of sub-segment lengths and the amount of sub-segment pairs covering each word. In general, the conclusion drawn from this experiment is that using either the training-free approach or a classification model trained on a corpus from a different domain are both valid options and they perform better than the alignment-based approach. Having a closer look to the data one can observe that the bigger differences are in the precision for "change" recommendations, where the MT-based approach outperforms the alignment-based approach.

7.5 Experiments on Reusability Across Machine Translation Systems

Table 16 presents the results of the experiments concerning MT system independence. Three models were trained on the three TM belonging to domain 02.40.10.40, but in each case

Θ(%)	training corpus	alignment A (%)	alignment NC (%)	MT-based trained A (%)	MT-based training-free A (%)	MT-based NC (%)
	02.40.10.40	93.90±.16	6.09±.15	95.14±.14		
≥ 60	05.20.20.10	91.77±.18	9.63±.19	**93.34±.16**	**93.27±.16**	**5.13±.14**
	06.30.10.00	90.23±.20	11.67±.20	92.02±.18		
	02.40.10.40	94.32±.18	5.90±.18	95.63±.16		
≥ 70	05.20.20.10	92.68±.21	9.71±.23	**94.03±.19**	**94.14±.18**	**5.18±.17**
	06.30.10.00	91.60±.23	11.61±.25	92.74±.20		
	02.40.10.40	95.10±.20	5.37±.21	96.39±.17		
≥ 80	05.20.20.10	93.82±.23	9.20±.26	**95.56±.19**	95.29±.20	**5.53±.21**
	06.30.10.00	93.21±.24	11.05±.28	94.27±.22		
	02.40.10.40	95.34±.26	4.93±.26	96.90±.21		
≥ 90	05.20.20.10	94.61±.28	8.90±.34	**96.53±.23**	**96.61±.22**	**5.86±.28**
	06.30.10.00	94.20±.30	10.27±.37	96.31±.23		

Table 15: Accuracy (A) and fraction of words not covered (NC) obtained by the alignment-based recommender, the trainng-free MT-based recommender, and the trained MT-based recommender when translating Spanish segments into English in domain 02.40.10.40. The results were obtained after re-using recommendation and alignment models learned from the TMs belonging to the domains indicated in the second column. The results obtained with models trained on domain 02.40.10.40 are included (highlighted in gray) as a reference. All the MT systems available were used for both training and testing. Statistical significance tests were carried out, separately for accuracy and for the fraction of words not covered. Differences in the results which are statistically significant with $p \leq 0.05$ are highlighted in bold type. For most values of Θ, the difference between the accuracy of both MT-based approaches is not statistically significant, but their differences with the alignment-based approach are statistically significant with $p \leq 0.05$.

Θ(%)	A (%) trained			A (%) training-free
	Apertium	**Google Translate**	**Power Translator**	
≥ 60	92.94±.17	95.08±.14	91.12±.19	**93.36±.16**
≥ 70	93.41±.20	95.54±.16	93.02±.20	**94.28±.18**
≥ 80	**95.57±.19**	96.34±.18	94.78±.21	95.41±.20
≥ 90	96.65±.22	96.99±.21	96.47±.23	96.72±.22

Table 16: Accuracy (A) obtained by the trained MT-based recommender and the training-free MT-based recommender when translating Spanish segments into English in domain 02.40.10.40 and using Google Translate as the MT system for testing. For the trained approach, the results were obtained after re-using recommendation models learned from the same TM and language pairs but using the three different MT systems. The results obtained using a model trained on Google (column in gray) are included as an upper-bound, but are not included in the comparison. For every value of Θ the highest accuracy, with a statistically significant difference with $p \leq 0.05$ compared to the other values, is highlighted in bold type.

using one of the three MT systems available. The models were used to translate segments in Spanish into English within the same domain and using Google Translate as the MT system, in order to obtain the sub-segment translations during testing. The results in this table are similar to those presented in the last set of experiments, in which the reusability across different domains was studied. In general terms, it would appear that the drop in accuracy when making "change" recommendations is quite similar for the models trained on Apertium and Power Translator. In addition, we observed that the accuracy obtained for these two models is similar to that obtained by the training-free recommender. The training-free approach is, in fact, that which performs best for $\Theta \geq 60\%$ and $\Theta \geq 70\%$ and the difference in accuracy is statistically significant with $p \leq 0.05$. However, for $\Theta \geq 80\%$ the trained approach using a model trained with Apertium is that which performs best. Finally, there is no difference among the results for $\Theta = 90\%$.

In general, it would appear that re-using models trained on an MT system which is different to that used for translation is feasible, although using the training-free approach can provide better results.

7.6 Error Analysis

The following is a sample of the most frequent errors made by the different approaches proposed in this work for word-keeping recommendation. The objective of this error analysis is to propose strategies to deal with these errors (when possible) in future work.

7.6.1 Errors Caused by Synonyms or Equivalent Expressions

Some of the incorrect change recommendations in our experiments resulted from the use of different synonyms in the translation proposal T and the reference T' used as a gold standard. Let us suppose a translation proposal (S, T)= (*"the natural isotopic abundance* *of lithium-6 is approximately 6,5 weight per cent (7,5 atom per cent).", "la proporción natural del isótopo litio-6 es de aproximadamente 6,5% del peso (7,5% de átomos).")*, for the sentence S'= *"the natural isotopic abundance of lithium-6 is approximately 6,5 weight % (7,5 atom %)."* whose reference translation is T'= *"la proporción natural del isótopo 6 en el litio es de aproximadamente 6,5% en peso (7,5% de átomos).".* As can be seen, S and S' are semantically equivalent and are written almost the same, although the percentage symbol (%) is used in S', while the expression *per cent* is used in S. Although these two options are equivalent, *per cent* is not considered to be part of the matching between S and S' in any of the occurrences. A sub-segment pair (σ, τ) with $\sigma =$ *"per cent"* and $\tau =$ *"%"*, may have led the two occurrences of the symbol % in T to be changed, when this is obviously not necessary. Since the symbol % was also used in the reference translation T', these are in fact considered to be wrong recommendations in the evaluation.

7.6.2 Errors Caused by Morphological Differences Between the Languages

This problem is in some respects similar to the previous one, although here the problem does not concern using different words for the same concept, but rather the presence of a word in one of the languages that may have different morphologies in the other language. For example, we found the proposal (S, T)= *("optical equipment as follows:","equipo óptico según se indica:")*, for the sentence S'= *"optical detectors, as follows:"* whose reference translation is T'= *"detectores ópticos según se indica:".* In this case the word *optical* is matched in both S and S', being singular in S and plural in S'. While in English both forms share the same orthography, in Spanish, the plural mark is added in T' (*ópticos*), therefore differing from the singular form in T (*óptico*). As a result, the word *óptico* would be probably recommended to be "kept", although it is not actually in the final translation (or at least not inflected in this way). It is worth noting that this would not be such a bad recommendation, since the difference between the word to be kept and the word needed for the translation is the same, but inflected in a different way. Whatever the case might be, it would be necessary to indicate this situations in some way, in order to let the user know that a change must be made.

7.6.3 Errors Caused by Fertility

This refers to the fact that the translation of a single word in one language is translated by two or more word in the other language. These words form a multi-word expression that can be translated properly only when using a sub-segment covering the whole expression. Sub-segments covering only a part of the expression can lead to out-of-context translations that produce wrong evidence. For instance, in the TU (S, T)= *("wavelength of less than 1400nm", "longitud de onda inferior a 1400nm")*, proposed for the sentence S'= *"wavelength equal to or greater than 1400nm"* whose reference translation is T'= *"longitud de onda igual o superior a 1400nm"* the word *wavelength* in English is translated as *longitud de onda* in Spanish. Since *wavelength* appears in both S and S', it is obvious that the three words in this multi-word expression should be kept. However, out of this context, word *de* can be translated as *of*, which also appears in S and is not matched in S'. The word *de* may therefore be obtaining "keeping evidence" from the sub-segments covering the whole expression *longitud de onda*, but "changing evidence" from the sub-segments covering only a part of the expression. This situation may easily result in a change recommendation. This is probably the most difficult error to fix, since it is motivated by the specifics of each language and may, in some cases, be extremely complex.

8. Concluding Remarks

In this paper we have presented a new approach to assist CAT users who employ TMs by providing them with recommendations as to which target-side words in the translation proposals provided by the CAT system have to be changed (modified or removed) or kept unedited. The method we propose imposes no constraints on the type of MT system to be used, since it is used as a black box. This method may use more than one MT system at the same time to obtain a set of features that are then combined using a binary classifier to determine whether target words have to be changed or should remain unedited. In any event, MT results are never presented to the translator. A version of this method which does not require any training is also proposed as an alternative.

The experiments carried out bear witness to the feasibility of the methods proposed by comparing the accuracy and coverage to those of a previous approach based on statistical word alignment for word-keeping recommendation. These experiments tackle the problem in different scenarios, comparing the results obtained for different domains, MT systems and language pairs. The results obtained confirm the viability of the MT-based methods proposed in this work, particularly in the case of the trained MT-based approach (see Section 4), which obtains better results as regards accuracy than those obtained with the statistical alignment-based approach (see Section 3). In the case of coverage, the results obtained with the MT-based approaches are in general worse than those obtained using the alignment-based approach when the alignment models are trained on in-domain TMs, but better when they are trained on out-of-domain TMs. The results also show a reasonable degree of independence of the MT-based models with respect to the domain of the TM and the MT system(s) used for training. These results suggest that there is no need to re-train the classifier for each new TM and, even more importantly, that it is not necessary to do so every time a new TU is added to the TM.

In general, the models trained for the MT-based recommender are much more portable across domains than those trained for the alignment-based approach. These approaches were compared to a training-free approach (see Section 5), which also uses MT as a source of evidence for word-keeping recommendation, but which does not need any training. The experiments confirm that the results obtained with the training-free MT-based recommender are worse than those obtained with the trained recommender when it is trained on the same TM to be used for translating new texts. However, it is advisable to use the training-free MT-based approach when no recommendation models for the same TM are available for the trained MT-based recommender.

In summary, the MT-based approaches perform better than the alignment-based approaches when accuracy is more important than coverage, or when they are trained on out-of-domain TMs. With regard to the MT-based approaches, it is better to use a trained MT-based recommender when a model is available for the pair of languages and MT system(s) to be used for translating, and to use the training-free MT-based recommender otherwise.

The principal conclusion of this work is that the three approaches are comparable and useful depending on the needs of the translator and the resources available for translation. It might even be possible to combine the three approaches (trained MT-based, training-free MT-based, and statistical alignment-based) in order to prove, for example, recommendations for those words not covered by some of the approaches and not by the others.

The results obtained in this study have also opened up other new horizons for future work, such as: extending the method so as to be able to not only provide the user with recommendations as to which words to keep unedited, but also actively suggest a translation for the words to change; trying alternative parametric classifiers; and using other sources of bilingual knowledge, such as glossaries, dictionaries, bilingual concordancers, etc. to improve the results of the MT-based approaches for word-keeping recommendation.

Appendix A presents a study that confirms the usefulness of word-keeping recommendation for translators, showing an improvement in productivity of up to 14% in a translation task from Spanish into English. We plan to extend these experiments to explore new ways of performing word-keeping recommendation. For instance, it would be interesting to compare the productivity of translators when receiving recommendations only for content words, optionally with partial recommendations (on stems), or receiving only change recommendations. We also believe that it would be interesting to evaluate the amount of minimum recommendations needed in a segment to make this tool useful for the translators, by computing the productivity of translators as regards proposals with low coverage. One of our main interests is to be able to model the cost of errors in recommendations, i.e. to confirm whether a wrong "keeping" recommendation is more expensive for a translator than a wrong "changing" recommendation. All these ideas require a new set of experiments with professional translators in order to obtain the optimal method with which to present recommendations, in order to maximise the improvement in productivity already shown in Appendix A.

A study on the impact of noise in the data set used for evaluation in this paper is included in Appendix B. This study uses an heuristic[31] to filter out free or wrong translations in the data sets. The translated materials obtained from the experiments described in Appendix A are additionally used as a clean data set produced directly from professional translators. The results in this appendix show that the accuracy in classification can be significantly improved when using clean data sets.

Finally, a prototype of a plug-in for the free/open-source CAT system OmegaT[32] which implements the training-free approach described in Section 5 as a proof of concept, is available and can be downloaded from `http://www.dlsi.ua.es/~mespla/edithints.html`. This prototype uses free on-line MT systems to perform word-keeping recommendation, thus confirming the technical feasibility of this approach as regards making on-the-fly recommendations in real-world settings.

31. This heuristic is based on the distance between the segment to translate S' and the source side of the translation proposal S, and the distance between the reference translation T' used as a gold standard for evaluation and the target side of the translation proposal T.
32. `http://www.omegat.org` [last visit: 15th May 2015]

Acknowledgments

This work is supported by the Spanish government through projects TIN2009-14009-C02-01 and TIN2012-32615. We would like to thank Juan Antonio Pérez-Ortiz, Andy Way and Harold Somers for their suggestions. We also thank the anonymous reviewers who helped to improve the original manuscript with their suggestions.

Appendix A. Experiment Concerning the Effect of Word-Keeping Recommendation on Translator Productivity

Word-keeping recommendation is a relatively new task. It is based on the assumption that providing translators using translation memory (TM) tools with hints about the words to change and to keep unedited in a translation proposal will increase their productivity. Although this might appear to be an obvious assumption, it needs to be empirically confirmed. The objective of the experiment described in this appendix is to verify the impact of word-keeping recommendation on translation productivity, independently of the approach used to obtain the recommendations.

A.1 Methodology

In this experiment, the productivity of professional translators was measured when translating several documents from English into Spanish by using the computer-aided translation (CAT) tool OmegaT, first without word-keeping recommendations and then with them. For this task, five translators with previous experience of using OmegaT were hired. Each of them had to translate three projects: a short training project (*training*), used only for familiarisation with the tool and the kind of documents to translate; a project to be translated with a standard version of OmegaT (*standard*); and a project to be translated with a modified version of OmegaT that provided word-keeping recommendations (*recommendation*).

The *training* project was the same for all five translators, while five different *standard* projects were created (one for each translator). The *standard* projects were reused as *recommendation* projects by rotating the translators, thus signifying that none of them translated the same project twice. The decision was made to use the translations obtained for the *standard* projects as the reference when computing the word-keeping recommendations for the *recommendation* projects; this would be equivalent to having a perfect classifier. This is often called an *oracle setting*.

Following the structure described, the experiment was driven in such a way that all the translations would be done at the same time, in the same room, and using identical computers. The experiment was divided into two phases: first, the *training* and the *standard* projects were translated, after which a break of about half an hour took place and the *recommendation* project was translated.

A.1.1 Corpus

The DGT-TM (Steinberger et al., 2012) translation memory published by the *European Commission Directorate-General for Translation* was used to build the translation memories and the translation projects used in this experiment. 90% of the document pairs in it were used as a TM after segment alignment. The remaining 10% was used to build the translation projects: documents were selected so that all their segments matched at least one translation unit (TU) in the TM with a fuzzy-match score (FMS) that was higher than or equal to 50%. Six translation projects were created from this selection: one containing a single document with 127 words (the *training* project) and five containing three different documents of about 1,000 words in total.

A.1.2 OmegaT

For the experiment, version 3.1 of the free/open-source CAT tool OmegaT was used with the plug-in *OmegaT SessionLog*[33] version 0.2, which silently logs all the actions performed by the translator. The initial version of OmegaT was modified to avoid exact matches (FMS=100%) being proposed, since it would not be possible to evaluate the impact of word-keeping recommendation on this kind of proposals.[34] A modified version of OmegaT was created that can also make word-keeping recommendations based on former translations. This version of the tool computes, for a given translation proposal, the edit distance between the reference translation and the proposal, and colours the words to be kept in green and the words to be removed from or replaced in the proposal in red. This means that the recommendations made by this version of OmegaT are the same as those that a professional translator would make when translating, i.e. perfect recommendations. (*oracle setting*).

A.2 Results

The results of the experiment are shown in Table 17. It is worth noting that it contains only the results for four of the data sets, given that one of the translators forgot to translate part of the *standard* project assigned to her, thus invalidating the corresponding results. This table shows the time devoted to translating each test set, both with and without using the word-keeping recommendation. Translation time was measured for each segment. The tool used to capture the edition information revealed that each segment is usually selected (or *visited*) several times during the whole process, both to translate it and to review it. In order to show this information more clearly, two different measures were obtained for each segment: total translation time (columns 2 and 3), which is the the time spent on a segment taking into account every visit to it, and edit time (columns 4 and 5), which is the time spent on translating it for the first time using a translation proposal. For this second measure, only the longest edit visit to each segment was taken into account, assuming that edits made during later visits corresponded to the review process. The last row of the table presents the total translation time for each column. As can be seen, the total time devoted to translation is reduced by more than 14% when using word-keeping recommendation. Moreover, editing time, on which word-keeping recommendation has the main impact, is reduced by more than 20%. This gain in translation time proved to be statistically significant with $p \leq 0.05$ when performing an *approximate randomisation test*

33. `https://github.com/mespla/OmegaT-SessionLog` [last visit: 15th May 2015]
34. It is assumed that an exact match provides a translation that does not need to be edited, and therefore, it is not possible to evaluate the advantage of word-keeping recommendations.

test set	total time		edition time	
	without WKR	with WKR	without WKR	with WKR
1	3,664s	2,611s	2,441s	1,917s
2	3,613s	3,467s	3,080s	2,293s
3	4,251s	3,709s	2,674s	2,310s
4	3,787s	3,315s	2,432s	1,937s
all test sets	15,315s	**13,102s**	10,627s	**8,457**s

Table 17: Time spent on translation. Columns 2 and 3 compare the total time spent translating each test set, respectively, using the version of OmegaT without and with word-keeping recommendation. Columns 4 and 5 present the same comparison, but only taking into account the time actually spent reviewing the test set.

with 1,000 iterations.[35] The free/open-source tool SIGF V.2 by Padó (2006) was used for these experiments.

The results obtained in this experiment confirm the assumption that word-keeping recommendation can significantly improve the productivity of translators who use translation memory tools. Although a more extensive experiment, including more translators and documents from other domains, would be needed to confirm this, current results are encouraging. In addition, all the translators participating in the experiment agreed that word-keeping recommendation is useful for translators when working with TM-based CAT tools.

It is worth noting that the experimental framework presented in this appendix has been specifically designed to measure word-keeping recommendation and the results obtained here cannot therefore be straightforwardly assumed for every translation project. For example, the projects translated in this experiment used a TM, thus ensuring that at least one translation proposal would be provided with an FMS that was higher than 50%, but a TM with this type of coverage may not be available for a given project. In addition, translations performed by humans were used in this experiment to compute the word-keeping recommendations, in what would usually be called a *gold standard*. These translations would obviously not be available in a real scenario and recommendations would be approximate. The use of gold-standard-based recommendations may also boost the confidence of the translators when using the tool, since in this experiment it was correct most of the times. We can therefore consider that these results correspond to an upper bound in productivity gain. Nevertheless, the results obtained in this experiment have allowed us to obtain a clearer idea of the usefulness of word-keeping recommendation and confirm the relevance of the problem of obtaining fast and accurate word-keeping recommendations.

35. Here, approximate randomisation is applied to the time devoted to translating each segment with and without word-keeping recommendation in the concatenated data sets. This method first computes the difference in time needed to translate the entire data sets. It then randomly interchanges the time spent translating some of these segments between both sets of results and recomputes this total time. If an equal or higher time gain can be obtained with these randomised lists of times, this means that the result is not significant.

Appendix B. Experiments with High Quality *Gold Standards*

In this appendix we tackle the problem associated with the use of free translations as references for evaluation. As already mentioned in Section 7, for a pair of segments (S'_l, T'_l) in our test set, we obtain all the matching TUs (S_i, T_i), and a set of word-keeping recommendations that are provided for every segment T_i. T'_l is then used as a gold standard for these recommendations for the purpose of evaluation. This method assumes that the way in which S'_l is translated into T'_l is similar to the way in which S_i is translated into T_i, thus enabling the use of T'_l as a reference.[36] However, this may not be the case for several reasons, such as wrong segment alignments, errors in translations, or, in our case, free (but still adequate) translations. Following the example shown in Section 4, we illustrate the impact of a free translation on our evaluation method. Let us assume that the segment S' to be translated is *"la situación política parece ser difícil"*, that a matching TU (S, T) retrieved by the CAT tool is *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult")*, and that the gold standard T' in the test set is *"the situation, from a political point of view, appears tortuous"*, which is a semantically valid translation of S', but very different to T. When checking the validity of the translation as occurred in Section 4, the only words common to T' and T are *the*, *appears*, and *situation*, and the remaining words would be considered as words to change in order to produce a correct translation. However, it is sufficient to replace the word *humanitarian* with *political* in T to produce a valid translation of S'. It is therefore obvious that T' is not a good reference with which to evaluate the recommendations performed on T.

As a consequence of the free translations in our test set, a fraction of the recommendations which are actually correct are considered inadequate during the evaluation since they do not match the reference in the test set. The accuracy obtained for all the approaches presented in this work is therefore lower than expected.

Although the loss of accuracy affects all the methods in this work in the same way and the conclusions that are obtained are therefore valid, we wished to see if more reliable results as regards the performance of these approaches could be attained. We therefore performed a set of additional experiments in order to bypass the problem of the free translations. On the one hand, we repeated some of the experiments shown in Section 7 but by using a constrained test set in which all those pairs of segments which were likely to be wrong (or free) translations were discarded. On the other hand, we performed an experiment by re-using the test set and the TMs described in Appendix A.

As mentioned previously, in the first group of experiments we defined a constraint in order to attempt to evaluate only those pairs of segments from the test set in Section 6.2 which are more reliable. This was done by employing a filtering based on the fuzzy-match score (FMS) used to choose the candidate TUs for a given segment to be translated. This condition relies on the assumption that the FMS between S and S' (FMS_S) should be similar to the FMS between T and T' (FMS_T), since the number of words that differed in both pairs of segments should be proportional for both languages. Based on this idea, we set a threshold ϕ so that only pairs of TUs fulfilling the condition $|\text{FMS}_S - \text{FMS}_T| \leq \phi$ were

36. By *similar* we mean that the matching parts between S'_l and S_i are translated in the same way, thus producing differences between T'_l and T_i only in those parts corresponding to the differences between S'_l and S_i.

$\Theta(\%)$	**domain**	**no filtering**		$\phi \leq 0.05$	
		$\mathrm{TU}_{\mathrm{avg}}$	N_{words}	$\mathrm{TU}_{\mathrm{avg}}$	N_{words}
≥ 60	02.40.10.40	3.71	95,881	1.59	44,240
	05.20.20.10	0.62	9,718	0.39	6,198
	06.30.10.00	5.68	34,339	0.52	6,956
≥ 70	02.40.10.40	2.36	65,865	1.16	31,022
	05.20.20.10	0.37	6,883	0.26	4,862
	06.30.10.00	0.99	10,327	0.26	4,194
≥ 80	02.40.10.40	1.58	46,519	0.97	24,928
	05.20.20.10	0.14	3,015	0.09	1,889
	06.30.10.00	0.45	4,726	0.09	2,143
≥ 90	02.40.10.40	0.70	26,625	0.50	15,154
	05.20.20.10	0.05	1,599	0.03	975
	06.30.10.00	0.03	1,268	0.03	943

Table 18: Average number of matching TUs ($\mathrm{TU}_{\mathrm{avg}}$) per segment and total number of words (N_{words}) for which to provide a recommendation when translating `es`→`en` for the three different domains. The results were obtained for different ranges of the FMS threshold (Θ), both when filtering with $\phi = 0.05$ and when no filtering was applied.

used for both training and testing. It is worth mentioning that some experiments were also performed by applying this filtering only to the test set, but the difference in the results was not significant. For our experiments, we arbitrarily set the value of ϕ to 0.05, i.e. a divergence of 5% was permitted between the FMS of the source language segments and that of the target language segments, since it is a threshold that constrains the examples used in a highly controlled scenario, but a reasonable number of samples is maintained for our experiments, as shown in Table 18.[37]

B.1 Experiments with Constrained Test Sets

This table shows, for the `es`→`en` language pair and for fuzzy-match scores in four different ranges, the average number of TUs matched per segment to be translated and the total number of words for which a recommendation should be provided. The results were obtained both when filtering with threshold ϕ and when no filtering was applied. It is worth noting that in the case of domains 05.20.20.10 and 06.30.10.00 there were noticeable differences in matching when no restriction was applied, while more similar data was obtained when filtering with $\phi = 0.05$. This has led us to believe that the TUs belonging to domain 05.20.20.10 are more regular in translation than those in domain 06.30.10.00. As will be

37. Note that the objective of these experiments is not to compare the different approaches (this has been already done), but rather to confirm whether an improvement in accuracy exists when using less noisy data sets. The statistical significance between the different approaches has not therefore been re-computed.

Θ(%)	Method	A (%)	NC (%)	P_{k} (%)	R_{k} (%)	P_{c} (%)	R_{c} (%)
≥ 60	**FMS-only**	84.7±.3	100%±0	84.7±.3	100%±0	—	0%±0
	alignment	96.4±.2	4.8±.2	98.2±.1	97.5±.2	85.8±.3	89.4±.3
	trained	96.9±.2	4.7±.2	97.9±.1	98.4±.1	90.8±.3	88.2±.3
	training-free	95.2±.2	4.7±.2	96.5±.2	97.9±.1	87.3±.3	79.9±.4
≥ 70	**FMS-only**	90.5±.3	100%±0	90.5±.3	100%±0	—	0%±0
	alignment	97.0±.2	4.7±.2	98.5±.1	98.2±.2	81.7±.4	83.9±.4
	trained	97.4±.2	4.4±.2	98.3±.2	98.8±.1	87.3±.4	82.7±.4
	training-free	96.1±.2	4.4±.2	97.1±.2	98.6±.1	83.4±.4	69.8±.5
≥ 80	**FMS-only**	93.2±.3	100%±0	93.2±.3	100%±0	—	0%±0
	alignment	97.4±.2	4.6±.3	98.7±.1	98.5±.2	77.1±.5	79.4±.5
	trained	98.0±.2	4.6±.3	98.7±.2	99.2±.1	86.5±.4	79.5±.5
	training-free	96.7±.2	4.6±.3	97.5±.2	98.0±.2	79.8±.5	61.7±.6
≥ 90	**FMS-only**	96.5±.3	100%±0	96.5±.3	100%±0	—	0%±0
	alignment	97.9±.2	4.1±.3	98.9±.2	98.9±.2	68.0±.8	68.0±.8
	trained	98.6±.2	4.2±.3	99.0±.2	99.6±.1	81.7±.6	62.8±.8
	training-free	98.3±.2	4.2±.3	98.9±.2	99.4±.1	74.0±.7	60.8±.8

Table 19: Comparison of the results obtained using the trained MT-based approach, the training-free MT-based approach, the alignment-based approach and the naïve FMS-only baseline. The accuracy (A) and fraction of words not covered (NC) are reported, together with the precision (P) and recall (R) for both keeping recommendations and changing recommendations. The results were obtained for several values of the FMS threshold Θ when translating es→en in domain 02.40.10.40, using all the MT systems available and filtering with $\phi = 0.05$ (see text).

observed, with this threshold, approximately half of the training samples are kept for domain 02.40.10.40 and about two thirds for domain 05.20.20.10. The case of domain 06.30.10.00 is different; the filtering removes far more training samples for low values of the FMS threshold Θ, while for higher values the loss is not so high, and similar to that of domain 05.20.20.10.

Table 19 is the equivalent of Table 12, which contains a detailed comparison of all the approaches, but using the filtering described above on the data set. As will be noted, the results obtained in this case are clearly better for all the approaches than those obtained in the experiments with no filtering.

Finally, Table 20 shows the accuracy obtained by both the trained MT-based approach and the alignment-based approach for all language pairs, as occurs in Table 10. It is worth noting that, although the differences between the results obtained with both approaches are similar, all of them are noticeably better.

B.2 Experiment With Human-Produced Test Sets

In this second group of experiments we used the documents described in Appendix A as a test set to evaluate word-keeping recommendation. In this case, the original documents in

	$\Theta \geq 60(\%)$		$\Theta \geq 70(\%)$		$\Theta \geq 80(\%)$		$\Theta \geq 90(\%)$	
lang. pair	trained	align-ment	trained	align-ment	trained	align-ment	trained	align-ment
es→en	96.9±.2	96.4±.2	97.5±.2	97.0±.2	98.0±.2	97.4±.2	98.5±.2	97.9±.2
en→es	95.1±.2	93.6±.2	96.4±.2	94.8±.2	97.1±.2	95.5±.2	97.8±.2	96.9±.3
de→en	96.9±.2	96.3±.2	97.7±.2	96.8±.2	98.3±.2	97.4±.2	98.3±.2	97.5±.3
en→de	96.3±.2	94.8±.2	97.2±.2	95.7±.2	97.6±.2	96.2±.2	97.9±.2	97.0±.3
fr→en	96.9±.2	96.7±.2	97.9±.2	97.7±.2	98.4±.2	98.0±.2	98.5±.2	98.1±.2
en→fr	95.9±.2	95.5±.2	97.4±.2	96.8±.2	98.1±.1	97.3±.2	98.3±.2	97.5±.2
fi→en	96.0±.2	96.0±.2	97.3±.2	97.1±.2	97.7±.2	97.5±.2	98.0±.3	97.5±.3
en→fi	96.3±.2	94.8±.3	97.2±.2	95.5±.3	97.7±.2	96.0±.3	97.7±.3	97.5±.3
es→fr	95.6±.2	95.3±.2	96.8±.2	96.5±.2	97.7±.2	97.2±.2	98.1±.2	97.4±.2
fr→es	95.2±.2	94.8±.2	96.7±.2	96.3±.2	97.3±.2	97.0±.2	97.4±.2	97.0±.3

Table 20: Accuracy (A) obtained with both the trained MT-based approach and the alignment-based approach when translating between all the language pairs in domain 02.40.10.40. The results were obtained for several values of the FMS threshold Θ and using all available MT systems for each language pair.

$\Theta(\%)$	A (%)	NC (%)
≥ 60	97.8±.1	10.0±.2
≥ 70	98.6±.1	8.5±.2
≥ 80	99.0±.1	8.1±.3
≥ 90	98.7±.2	7.3±.4

$\Theta(\%)$	A (%)	NC (%)
≥ 60	95.6±.1	9.8±.2
≥ 70	96.2±.1	8.5±.2
≥ 80	96.7±.2	8.1±.2
≥ 90	96.4±.2	8.1±.3

Table 21: Accuracy (A) and fraction of words not covered (NC) obtained when translating with the trained MT-based approach by reusing the data set described in Appendix A. The left-hand table contains the results when translating Spanish into English, while the right-hand table contains the results when translating English into Spanish.

Spanish were translated into English by professional translators, who were told to translate them as faithfully as possible. These parallel documents were therefore expected to totally fit the requirements of the evaluation.

In this experiment, the TM used by the professional translators in Appendix A was used to evaluate the translation of the texts from English into Spanish and vice versa. In-domain models were also trained on this TM which, as already mentioned above, consists of only 629 TUs. Table 21 presents the accuracy and the fraction of words not covered that were obtained for this data set, for en→es and for es→en. Although the coverage is slightly lower than that obtained by the system with other data sets, the accuracy is much better, and is even better than that obtained with the constrained test sets.

The results presented in this appendix allow us to confirm that the accuracy of the approaches presented in this work may be noticeably higher than those presented in Section 7, but the lack of a valid gold standard for our experiments only allows us to approximate these results.

References

Ahrenberg, L., Andersson, M., & Merkel, M. (2000). *Parallel text processing: alignment and use of translation corpora*, chap. A knowledge-lite approach to word alignment. Kluwer Academic Publishers. Edited by J. Véronis.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems 24*, pp. 2546–2554. Curran Associates, Inc.

Bertoldi, N., Farajian, A., & Federico, M. (2009). Online word alignment for online adaptive machine translation. In *Proceedings of the Workshop on Humans and Computer-assisted Translation*, pp. 120–127, Gothenburg, Sweden.

Biçici, E., & Dymetman, M. (2008). Dynamic translation memory: Using statistical machine translation to improve translation memory fuzzy matches. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics*, Vol. 4919 of *LNCS*, pp. 454–465, Haifa, Israel.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., & Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, USA.

Bourdaillet, J., Huet, S., Langlais, P., & Lapalme, G. (2010). TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation*, *24*(3–4), 241–271.

Bowker, L. (2002). *Computer-aided translation technology: a practical introduction*, chap. Translation-memory systems, pp. 92–127. University of Ottawa Press.

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

de Gispert, A., Blackwood, G., Iglesias, G., & Byrne, W. (2013). N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, *27*(2), 85–114.

Depraetere, I. (2008). LEC Power Translator 12. *MultiLingual*, *September 2008*, 18–22.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification* (second edition). John Wiley and Sons Inc.

Esplà, M., Sánchez-Martínez, F., & Forcada, M. L. (2011). Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. In *Proceedings of the 15th Annual Conference of the European Associtation for Machine Translation*, pp. 81–89, Leuven, Belgium.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2011). Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the 13th Machine Translation Summit*, pp. 172–179, Xiamen, China.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2015). Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Associtation for Machine Translation*, pp. 19–26, Antalya, Turkey.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2012a). Using external sources of bilingual information for on-the-fly word alignment. Tech. rep., Universitat d'Alacant.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2012b). A simple approach to use bilingual information sources for word alignment. *Procesamiento de Lenguaje Natural*, *49*.

European Commission Directorate-General for Translation (2009). *Translation Tools and Workflow*. Directorate-General for Translation of the European Commission.

European Commission Joint Research Center (2007). EUR-Lex pre-processing. http://optima.jrc.it/Resources/Documents/DGT-TM_EUR-LEX-preprocessing.pdf. Last retrieved: 15th May 2015.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, *25*(2), 127–144. Special Issue on Free/Open-Source Machine Translation.

Foster, G., & Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pp. 128–135, Prague, Czech Republic.

Free Software Fundation (2007). GNU general public license, version 3. http://www.gnu.org/licenses/gpl.html. Last retrieved: 15th May 2015.

Gao, Q., Lewis, W., Quirk, C., & Hwang, M. (2011). Incremental training and intentional over-fitting of word alignment. In *Proceedings of the 13th Machine Translation Summit*, pp. 106–113, Xiamen, China.

Gao, Q., & Vogel, S. (2008). Parallel implementations of word alignment tool. In *Proceedings of the Software Engineering, Testing, and Quality Assurance for Natural Language Processing Workshop*, pp. 49–57, Columbus, USA.

Garcia, I. (2005). Long term memories: Trados and TM turn 20. *Journal of Specialised Translation*, *4*, 18–31.

Garcia, I. (2012). Machines, translations and memories: language transfer in the web browser. *Perspectives*, *20*(4), 451–461.

Gough, N., Way, A., & Hearne, M. (2002). Example-based machine translation via the web. In Richardson, S. D. (Ed.), *Machine Translation: From Research to Real Users*, Vol. 2499 of *Lecture Notes in Computer Science*, pp. 74–83. Springer Berlin Heidelberg.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, *11*(1), 10–18.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

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

Koehn, P., & Senellart, J. (2010). Convergence of translation memory and statistical machine translation. In *Proceedings of the 2nd Joint EM+/CNGL, Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*, pp. 21–31, Denver, USA.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., & Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, USA.

Kranias, L., & Samiotou, A. (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*, pp. 331–334, Lisbon, Portugal.

Kuhn, R., Goutte, C., Isabelle, P., & Simard, M. (2011). Method and system for using alignment means in matching translation. USA patent application: US20110093254 A1.

Lagoudaki, E. (2008). The value of machine translation for the professional translator. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, pp. 262–269, Waikiki, USA.

Langlais, P., & Simard, M. (2002). Merging example-based and statistical machine translation: An experiment. In Richardson, S. (Ed.), *Machine Translation: From Research to Real Users*, Vol. 2499 of *Lecture Notes in Computer Science*, pp. 104–113. Springer Berlin Heidelberg.

Läubli, S., Fishel, M., Volk, M., & Weibel, M. (2013). Combining statistical machine translation and translation memories with domain adaptation. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pp. 331–341, Oslo, Norway.

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

Ma, Y., He, Y., Way, A., & van Genabith, J. (2011). Consistent translation using discriminative learning: A translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 1239–1248, Portland, Oregon.

Marcu, D. (2001). Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, ACL '01, pp. 386–393, Toulouse, France.

Meyers, A., Kosaka, M., & Grishman, R. (1998). A multilingual procedure for dictionary-based sentence alignment. In *Machine translation and the information soup: Proceedings of the third conference of the Association for Machine Translation in the Americas*, Vol. 1529 of *LNCS*, pp. 187–198, Langhorne, USA.

Nießen, S., Vogel, S., Ney, H., & Tillmann, C. (1998). A DP based search algorithm for statistical machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pp. 960–967, Montreal, Canada.

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

Padó, S. (2006). *User's guide to* `sigf`*: Significance testing by approximate randomisation.*

Sánchez-Martínez, F., Carrasco, R. C., Martínez-Prieto, M. A., & Adiego, J. (2012). Generalized biwords for bitext compression and translation spotting. *Journal of Artificial Intelligence Research*, *43*, 389–418.

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*, pp. 65–72, Edmonton, Canada.

Simard, M., & Isabelle, P. (2009). Phrase-based machine translation in a computer-assisted translation environment. In *Proceedings of the 12th Machine Translation Summit*, pp. 120–127, Ottawa, Canada.

Simard, M., & Langlais, P. (2001). Sub-sentential exploitation of translation memories. In *Proceedings of the Machine Translation Summit VIII*, pp. 335–339, Santiago de Compostela, Spain.

Somers, H. (2003). *Computers and translation: a translator's guide*, chap. Translation memory systems, pp. 31–48. John Benjamins Publishing, Amsterdam, Netherlands.

Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, *14*(2), 113–157.

Specia, L., Raj, D., & Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, *24*(1), 39–50.

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

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC'12, pp. 454–459, Istambul, Turkey.

Ueffing, N., & Ney, H. (2005). Word-level confidence estimation for machine translation using phrase-based translation models. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 763–770, Vancouver, Canada.

Veronis, J., & Langlais, P. (2000). Evaluation of parallel text alignment systems. In Veronis, J. (Ed.), *Parallel Text Processing*, Vol. 13 of *Text, Speech and Language Technology*, pp. 369–388. Springer Netherlands.

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

Zhechev, V., & van Genabith, J. (2010). Seeding statistical machine translation with translation memory output through tree-based structural alignment. In *Proceedings of the COLING'10, Workshop on Syntax and Structure in Statistical Translation*, pp. 43–51, Beijing, China.