Automatic Translation Memory Fuzzy Match Post-Editing: A Step beyond Traditional TM/MT Integration

Lambros Kranias, Anna Samiotou

ESTeam AB
8 Sikelianou Str 14671, Athens, Greece
esteam@otenet.gr

Abstract

An innovative way of integrating Translation Memory (TM) and Machine Translation (MT) processing is presented which goes beyond the traditional "cascade" integration of Translation Memory and Machine Translation. The new method aims to automatically post-edit TM similar matches by the use of an MT module thus enhancing the TM fuzzy (similar) scores as well as enabling the utilisation of low-score TM fuzzy matches. This leads to substantial translation cost reduction.

The suggested method, which can be classified as an Example-Based Machine Translation application, is analysed and examples are provided for clarification. It is evaluated through test results that involve human interaction. The method has been implemented within the ESTeam Translator (ET) Language Toolbox and is already in use in the various commercial installations of ET.

1. Automatic Translation Memory Fuzzy Match Post-Editing

According to the standard TM paradigm (Nagao, 1984), an input text unit (usually a sentence) to be translated is matched against the source language part of translation pairs stored in the TM. If an identical (full) or similar (fuzzy) match is located, then the system suggests its target language equivalent as the translation of the original text unit and lets the user accept/edit this suggestion in order to correspond accurately to the translation of the input text unit. When no full/fuzzy match can be located, the option is usually offered to invoke MT processing to translate the input text unit. The method proposed in this paper, can be classified as an Example-Based Machine Translation application (Somers, 1999), taking the TM-MT integration one step further manipulating the fuzzy match result by invoking MT (**in context**) in order to automatically correct the TM-based translation suggestion.

We denote as S_{inp-SL} the input text unit, for example a sentence, consisting of words to be translated from the Source Language (SL) into the Target Language (TL). Suppose that the TM contains a text-unit pair, for example sentences again, denoted as S_{ref-SL} and S_{ref-TL}. The standard definition of a "fuzzy match translation" is that if S_{inp-SL} is similar to S_{ref-SL}, through the similarity of (some of) their words, then S_{ref-TL} is proposed as the translation of S_{inp-SL} (to be verified/edited by a human translator).

The suggested method exploits fuzzy match information $M(S_{inp-SL}, S_{ref-SL})$ as well as word-alignment information $A(S_{ref-SL}, S_{ref-TL})$ referring to the TM text-unit pair, in order to apply modifications on S_{ref-TL} to correspond to the translation of S_{inp-SL}. The fuzzy match information $M(S_{inp-SL}, S_{ref-SL})$ defines the "links" between words of S_{inp-SL} and S_{ref-SL}, in other words it defines which input-SL word has matched to which reference-SL word. This type of information is standard in all TM systems since it is used in order to estimate the similarity score of a match. The word-alignment information $A(S_{ref-SL}, S_{ref-TL})$, however, is anything but standard. The bottleneck of the application of "Fuzzy Match Post Editing" is the existence of word-alignment information (for the TM contents), which enables the appropriate correction of the TL reference text units. Word-alignment information defines the translation "links" between words of reference-SL and reference-TL text units (the TM pair), in other words it defines which word/phrase of the S_{ref-SL} translates to which word/phrase of the S_{ref-TL} (and can, in general, include phrases with non-consecutive words). This information, which is not necessarily exhaustive, can be either calculated on-line (by looking up an MT dictionary) or can be pre-stored in the TM. In the ESTeam Translator system, word-alignment information is available, through a process of automatically aligning text units at various text levels (paragraphs, sentences, subsentences) (Meyers 1998, Ahrenberg et al, 2000) by the use of (among other resources) an MT Dictionary of words and phrases. The MT Dictionary defines the relevance of two text units being compared (by defining translation links between their words) and then "marks" the corresponding word-alignment information to be later used for the application of "Fuzzy Match Post Editing".

The basic idea of the "Fuzzy Match Post Editing" is quite simple and it is graphically depicted in Figure 1 for the case of an example involving all supported actions:

Insertion(s) of Word(s)

It identifies **mismatched** words in S_{inp-SL} and based on the fuzzy match information $M(S_{inp-SL}, S_{ref-SL})$, which provides "anchor points" in the vicinity of these mismatched words, it tries to identify the corresponding "missing word" positions in S_{ref-SL}. It then searches in $A(S_{ref-SL}, S_{ref-TL})$ for potential available word-alignment

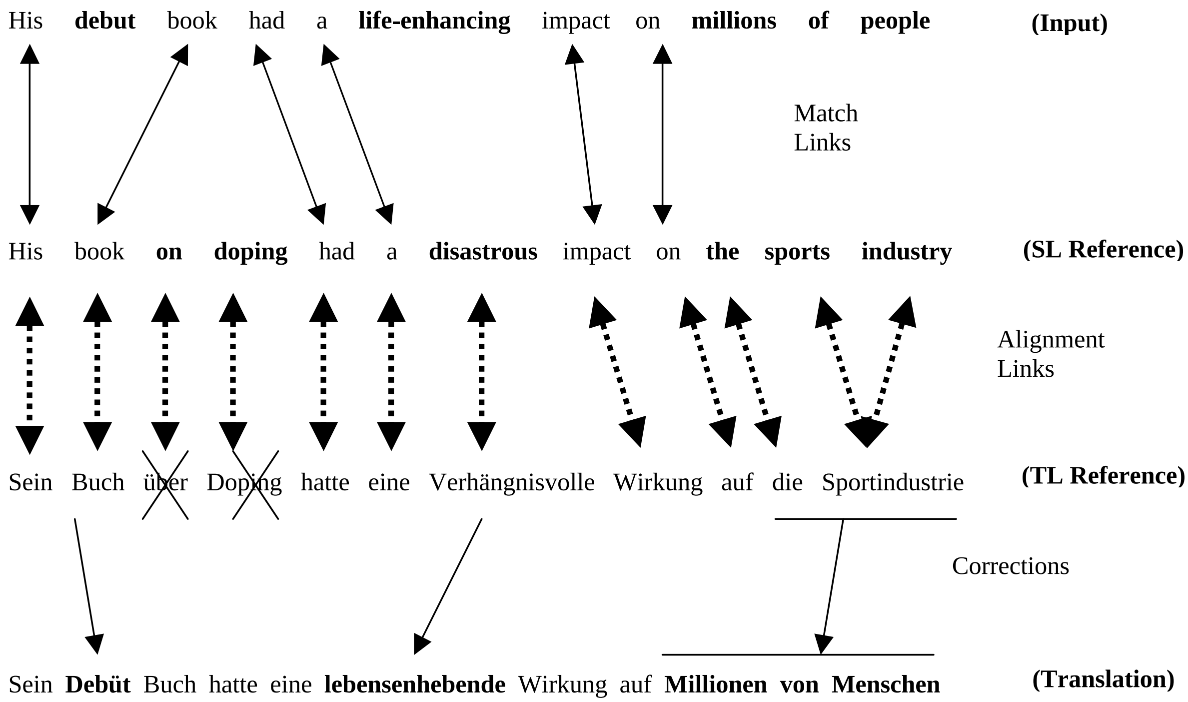

Figure 1: An Example
(mismatched and "corrected" words appear in boldface)

links for the vicinity of the identified "missing word" positions in $S_{ref\text{-}SL}$. If such word-alignment links exist, and if they point to words in $S_{ref\text{-}TL}$ that retain the same order (as in $S_{inp\text{-}SL}$), then it invokes MT in order to translate the identified mismatched words in $S_{inp\text{-}SL}$ and places the translation in the appropriate position in $S_{ref\text{-}TL}$.

Deletion(s) of Word(s)

It identifies **mismatched** words in $S_{ref\text{-}SL}$ and based on the fuzzy match information $M(S_{inp\text{-}SL}, S_{ref\text{-}SL})$, which provides "anchor points" in the vicinity of these mismatched words, it tries to verify that the mismatched words are indeed extra words appearing in $S_{ref\text{-}SL}$. It then searches in $A(S_{ref\text{-}SL}, S_{ref\text{-}TL})$ for potential available word-alignment links for these mismatched (extra) words. If such word-alignment links exist, then it deletes the corresponding words in $S_{ref\text{-}TL}$.

Replacement(s) of Word(s)

It identifies **mismatched** words in $S_{inp\text{-}SL}$ and based on the fuzzy match information $M(S_{inp\text{-}SL}, S_{ref\text{-}SL})$, which provides "anchor points" in the vicinity of these mismatched words, it tries to identify the corresponding **mismatched** words in $S_{ref\text{-}SL}$. It then searches in $A(S_{ref\text{-}SL}, S_{ref\text{-}TL})$ for potential available word-alignment links for the identified mismatched words and their vicinity in $S_{ref\text{-}SL}$. If such word-alignment links exist, it invokes MT in order to translate the identified mismatched words in $S_{inp\text{-}SL}$ and then it replaces the translation(s) in the appropriate position in $S_{ref\text{-}TL}$.

Each correction that is applied leads to a re-evaluation (increase) of the fuzzy match score since it simulates handled mismatches as normal matches.

An example is provided for clarification, which demonstrates word insertion, deletion and replacement (Figure 1).

The fuzzy match score is initially calculated as 52%. The proposed method locates the mismatched words (input and reference EN sentences):

- "**debut**" ($S_{inp\text{-}SL}$)
  Bounded by the fully matched word sequence "His … book", it is identified as an extra $S_{inp\text{-}SL}$ word associated to the position between the first and second $S_{ref\text{-}SL}$ words
- "**on doping**" ($S_{ref\text{-}SL}$)
  Bounded by the fully matched word sequence "book … had a", it is identified as an extra $S_{ref\text{-}SL}$ word
- "**life-enhancing**" ($S_{inp\text{-}SL}$)
  Bounded by the fully matched word sequence "had a … impact", it is identified as a mismatched $S_{inp\text{-}SL}$ word associated to the word "**disastrous**" in $S_{ref\text{-}SL}$.
- "**on millions of people**" ($S_{inp\text{-}SL}$)
  Bounded by the fully matched word sequence "impact on … ", it is identified as a mismatched

$S_{inp\text{-}SL}$ word sequence associated to the word sequence "**the sports industry**" in $S_{ref\text{-}SL}$.

It then tries to identify the equivalents of the "mismatched parts" in $S_{ref\text{-}SL}$ by looking up the (pre-existing) word alignment information (depicted in Figure 1). It results in the following modifications:

- **Insertion**
  The word "**Debüt**" is inserted (as the machine translation of "**debut**") between the words "**Sein**" and "**Buch**" in $S_{ref\text{-}TL}$.

- **Deletion**
  The word sequence "**über Doping**" is deleted in $S_{ref\text{-}TL}$ (since it is linked to the mismatched word sequence in $S_{ref\text{-}SL}$ "**on doping**")

- **Replacement**
  The word "**Verhängnisvolle**" is replaced by "**lebensenhebende**" (the machine translation of the phrase "**life-enhancing**") in $S_{ref\text{-}TL}$.

- **Replacement**
  The word sequence "**die Sportindustrie**" is replaced by "**Millionen von Menschen**" (as the machine translation of the phrase "**millions of people**") in $S_{ref\text{-}TL}$.

Thus producing the final translation:

"Sein Debüt Buch hatte eine lebensenhebende Wirkung auf Millionen von Menschen"

which is a correct translation of the input sentence. The re-estimated fuzzy match score is now 90% (even though all mismatched parts were actually handled, the score includes a "penalty" factor due to the use of machine translation which is not guaranteed to be accurate).

2. Evaluation

The proposed method for automatic fuzzy match post-editing is evaluated through a large scale experiment: A big translation batch is formed, consisting of 20,000 English sentences to be translated in German and French. The translation batch is automatically pre-translated, using the ESTeam TM, setting as minimum acceptable fuzzy match score the value of 70%.

The application of automatic fuzzy match post-editing increases the fuzzy score of the corresponding sentences. So, in the test we lower the minimum acceptable fuzzy match score to the value of 50% expecting that some of the fuzzy matches in the score region 50-70% will be automatically post-edited into matches of score 70%.

Human experts are only presented with those fuzzy matches with score higher than 70% which were automatically post-edited. Their task is, on one hand, to verify whether the automatic post-editing is successful (since, in general, the automatic post-editing might produce improper corrections to the target language reference sentence) and, on the other, to evaluate whether the corresponding fuzzy score increase is accurate. The results are presented in Table 3.
The error rates reported in table 3 are higher for the translation direction English to French. This can be explained by the fact that, unlike German, French follows a different general word order than French. This complicates the positioning of corrected word items in the $S_{ref\text{-}TL}$, resulting to an increased number of improper application of the automatic fuzzy match post-editing.

The increase of the fuzzy match scores leads to translation cost reduction. This can be calculated depending on the translation cost schema applied in relation to fuzzy match scores. A typical translation cost schema is:

- full matches cost zero
- fuzzy matches in the score region 90-99% cost 20% of the normal cost
- fuzzy matches in the score region 80-89% cost 40% of the normal cost
- fuzzy matches in the score region 70-79% cost 60% of the normal cost

According to this schema, the application of automatic fuzzy match post-editing in this experiment lead to a translation cost reduction of about 8% for the translation direction English-German and about 6% for the translation direction English-French, which are definitely important figures for large scale translation projects.

3. Conclusion

The results presented in this paper prove that the use of Automatic Fuzzy Match Post-Editing can lead to significant translation cost reduction. The results can be even more favourable (as compared to the results presented in this paper) if the method is configured to operate in a less strict way, processing insertions / deletions / replacements based on the left **or** right context of the text unit under investigation. However, in that case, the rate of "inappropriate" results of the method would also increase.

References

- Ahrenberg L., Andersson M. & Merkel M. (2000), A knowledge-lite approach to Word Alignment. In Veronis, J., "Parallel Text Processing: Alignment and Use of Translation Corpora." (Kluwer Academic, 2000).
- Meyers A., Kosaka M. and Grishman R. (1998). "A Multilingual Procedure for Dictionary-Based Sentence Alignment", Proceedings of ACL-COLING-98, Montreal, Canada.
- Nagao M. (1984) A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. Artificial an' Human Intelligence, ed. Elithorn A. and Banerji R., North-Holland, pp 173-180.
- Somers H. (1999) Review Article: Example-Based Machine Translation. Machine Translation, 14(2), 113-157

	EN-DE	**EN-FR**
Number of Sentences	20,000	20,000
Number of full/fuzzy matches for which **no** post-editing is possible	4,955	5,312
Number of fuzzy matches above 70% which were automatically post-edited	1,044	986
Number of errors in the category above	162	240
Average fuzzy match score **increase** due to **correct** application of post-editing	18%	15%
Number of fuzzy matches between 50-70% which were automatically post-edited resulting in fuzzy scores above 70%	1367	1128
Number of errors in the category above	226	305
Average fuzzy match score due to **correct** application of post-editing in the score region 50-70%	81%	78%

Table 3: Evaluation Results English-French