Missed opportunities in translation memory matching

Friedel Wolff[1], Laurette Pretorius[1], Paul Buitelaar[2,1]
[1]College of Graduate Studies
University of South Africa
[2]INSIGHT Center for Data Analytics
National University of Ireland, Galway
wolfff@unisa.ac.za, pretol@unisa.ac.za, paul.buitelaar@deri.org

Abstract

A translation memory system stores a data set of source-target pairs of translations. It attempts to respond to a query in the source language with a useful target text from the data set to assist a human translator. Such systems estimate the usefulness of a target text suggestion according to the similarity of its associated source text to the source text query. This study analyses two data sets in two language pairs each to find highly similar target texts, which would be useful mutual suggestions. We further investigate which of these useful suggestions can not be selected through source text similarity, and we do a thorough analysis of these cases to categorise and quantify them. This analysis provides insight into areas where the recall of translation memory systems can be improved. Specifically, source texts with an omission, and semantically very similar source texts are some of the more frequent cases with useful target text suggestions that are not selected with the baseline approach of simple edit distance between the source texts.

Keywords: Translation memory, text similarity, edit distance

1. Introduction

A translation memory (TM) is a repository in which the user can store previously translated (target) texts paired with the corresponding source text in a structured way. TMs are widely used in the translation industry (Lagoudaki, 2006) and have been shown to improve translator productivity and consistency (Morado Vázquez, 2012). A TM system functions as an information retrieval system that tries to retrieve one or more suggestions from a TM database that would assist the translator in his/her current translation task. In contrast with machine translation systems which can always provide some suggestion, a TM system is likely to only provide suggestions for a subset of queries. O'Brien (2007) indicated that the cognitive load experienced by the translator is highest for segments with no suggestions, compared to segments with either TM suggestions or suggestions from machine translation. Improving recall for a TM system is therefore desirable. This study focusses on those suggestions that would be useful, but are not selected by a baseline similarity measure.

A string similarity measure such as the Levenshtein distance (Levenshtein, 1966) is commonly used to measure the similarity of source texts in TM systems. In this study, we specifically use the four operation edit similarity over characters as the baseline. The four operations in this case are insertion, deletion, substitution and identity (no change).

In normal operation a TM system searches in translation pairs for source text similar to the current segment for translation. If found, the associated target texts are presented to aid the translator in translation of the current segment. A useful suggestion needs little or no editing to transform it into the desired translation in the target text. Other target segments in the TM might exist that could be equally useful, but are not selected because their source texts are not sufficiently similar to the query (current segment for translation).

Since we want to improve TM systems by additionally selecting these useful target segments with highly dissimilar source text, we pose these research questions: What are the attributes of these source texts that prohibit them from being selected using the baseline approach, and, if selected, how significant an improvement might we expect? These source-target pairs are what we refer to as "missed opportunities". In addressing these questions, we structure the paper as follows: In the next section we outline the approach that we follow to identify "missed opportunities". Since thresholds are important we briefly discuss the aspect of parameters in section 3. In section 4 we introduce the datasets that we use in our study. Section 5 constitutes the main and novel part of the paper, namely the analysis of the identified missed opportunities in order to identify common patterns that could be exploited in categorising such opportunities towards improving a TM. Possible categories are found and discussed in some detail. The paper is concluded with preliminary findings, possible solution approaches and suggestions for future work.

2. Approach

We proceed as follows: For all (s_i, t_i) pairs in a given TM, identify all (s_j, t_j) pairs with target segment t_j highly similar to t_i, i.e. $similarity(t_i, t_j) \geq 95\%$. From all identified (s_j, t_j) pairs, select those for which the source segment similarity with s_i is below the required similarity threshold, i.e. $similarity(s_i, s_j) < 65\%$. These remaining (s_j, t_j) pairs represent our "missed opportunities".

We then manually investigate these (s_i, s_j) pairs to understand what the attributes are that resulted in these not being selected using the baseline approach. In each case the two source texts are compared to see if there is an apparent relation between them that would reveal why their associated target texts might be similar. These relations are categorised, and the categories and the number of times they were identified give us an indication of where we need to focus our

efforts to improve recall in a TM system. This could also serve as a test set for developing similarity measures that address these identified shortcomings.

In this approach, no linguistic knowledge of the chosen target languages is used. This avoids problems of incomplete knowledge causing bias one way or the other. Applying knowledge of the target text could be useful in certain cases, for example to identify polysemous words/phrases in the target text to explain why their corresponding source texts differ markedly. However, this could become a slippery slope where classifications rely heavily on the language knowledge of individual judges, and results could unnecessarily differ between languages. The preliminary nature of this study, and the desire to investigate linguistically unrelated target languages, also suggests a zero-knowledge approach (in terms of the target language). As will be apparent later, knowledge of the source text (English) was used to identify relationships between the identified pairs of source text.

3. Parameters

The ideal thresholds to be used for determining similarity or relevance in a TM system are subjective, and often configurable in tools for computer assisted translation. For the purposes of this study, we used 95% as the similarity threshold for identifying highly similar target texts, and only considered opportunities "missed" where the source texts had a similarity below 65%.

A target text similarity of 95% can be considered a *very* useful suggestion, needing only a small amount of editing work to transform it into the reference translation. The results obtained by O'Brien (2007) suggest that fuzzy suggestions in the fuzzy match value range of 91% to 99% result in the least cognitive effort to process, and translators attain a higher speed, compared to suggestions at lower fuzzy match values. In this study, we therefore put the bar even higher—accepting only target texts with similarity of 95% or higher.

The threshold of 65% for source text similarity is chosen because existing translation tools are likely to exclude all suggestions with similarity below 70% if left at their default settings (SDL plc., 2013; Andre, 2013). It should be noted, however, that the matching methods for translation tools are often proprietary, and that the numbers themselves might not be directly comparable to the character based Levenshtein similarity used in this study. Several authors have confirmed that the measures in use by proprietary tools are character based similarity measures (Macklovitch and Russell, 2000; Somers, 2003).

This choice for thresholds therefore attempts to identify target texts which are very likely to be useful to a translator, and source texts that are very unlikely to be chosen by translation tools at their default settings.

4. Resources

Two different translation memories were investigated, both in two language pairs. The DGT-TM Release 2011 (Steinberger et al., 2012) contains translations of legislative texts of the European Union. We specifically used the 2004_1 subset. This corpus is published as a sentence aligned parallel corpus. The alignment was done automatically, and this corpus was used with the alignments as published.

The GNOME corpus contains the translations of the user interfaces of the GNOME 3.8 desktop environment.[1] The segments in the GNOME corpus are much shorter on average, can include markup, placeholders and other non-textual elements as is common in software localisation. Segments frequently contain very short strings, but can also contain complete paragraphs with multiple sentences. This corpus is not sentence aligned. In all cases English was used (or assumed) as source language. French and Hungarian were chosen as two linguistically unrelated target languages that occur in both datasets with a similar amount of text. See table 1 for an overview of the corpus statistics.

Corpus	Segments	% unique	Avg. words
DGT (fr)	102386	69%	16.6
DGT (hu)	71616	64%	15.8
GNOME (fr)	40801	89%	4.3
GNOME (hu)	40801	88%	4.3

Table 1: Corpus statistics

5. Classification

The main goal of this study is to identify common patterns among the "missed opportunities" to be able to improve TM matching. Table 2 shows the main categories that were identified, as well as their frequencies per corpus.

The classification into these categories is to some extent subjective. For example, whether or not an omission is a translation error or a warranted stylistic choice might be a matter of opinion to be evaluated in each case in the original context. The original context (a document, or application GUI) is considered not to be available for this study since we are dealing with translation memories and not the final products of translation.

Several translation pairs returned from the investigation had no obvious reason for having very dissimilar source text and highly similar target text. These could be due to translation errors, quality issues in the source text, or any number of issues arising out of the local context, which is not accessible from this investigation. The cases with no obvious reason were ignored in the rest of the investigation.

In longer segments, there were often more than one difference between the two source texts. In such cases the source of the biggest diversion in terms of the edit distance was classified.

By following the approach in section 2 we were able to identify four main categories according to which missed opportunities could be further exploited, namely semantics, omission, text normalisation and misalignment. We discuss each category in subsequent subsections.

5.1. Semantics

Some of the missed opportunities identified can roughly be grouped together as relating to semantics. These include

[1] Available from https://l10n.gnome.org/releases/gnome-3-8/

Classification	Corpus	Segments
No classification	DGT (fr)	581
Synonyms	DGT (fr)	68
Paraphrase	DGT (fr)	76
Active/passive	DGT (fr)	1
Word order	DGT (fr)	62
Abbreviation	DGT (fr)	6
Omission	DGT (fr)	117
Case	DGT (fr)	22
Normalisation	DGT (fr)	0
Misalignment	DGT (fr)	59
No classification	DGT (hu)	532
Synonyms	DGT (hu)	43
Paraphrase	DGT (hu)	29
Active/passive	DGT (hu)	0
Word order	DGT (hu)	22
Abbreviation	DGT (hu)	6
Omission	DGT (hu)	66
Case	DGT (hu)	2
Normalisation	DGT (hu)	0
Misalignment	DGT (hu)	142
No classification	GNOME (fr)	405
Synonyms	GNOME (fr)	180
Paraphrase	GNOME (fr)	34
Active/passive	GNOME (fr)	6
Word order	GNOME (fr)	34
Abbreviation	GNOME (fr)	19
Omission	GNOME (fr)	55
Case	GNOME (fr)	3
Normalisation	GNOME (fr)	10
Misalignment	GNOME (fr)	0
No classification	GNOME (hu)	504
Synonyms	GNOME (hu)	184
Paraphrase	GNOME (hu)	31
Active/passive	GNOME (hu)	7
Word order	GNOME (hu)	24
Abbreviation	GNOME (hu)	22
Omission	GNOME (hu)	92
Case	GNOME (hu)	2
Normalisation	GNOME (hu)	24
Misalignment	GNOME (hu)	0

Table 2: Classification statistics

synonymy, paraphrases, including active/passive variation, as well as abbreviations. In these cases the two source texts represent the same or very similar meaning, even if their character similarities are lower than the threshold. Some examples are shown in table 3.

The most significant of these categories is **synonymy** in the source text. An **abbreviation** can be considered semantically equivalent to its full form, or similar to a synonym, but was classified separately. Because of the length difference (in characters) between abbreviations and the full forms they represent, length based similarity measures (such as Levenshtein distance) are likely to be unable to identify corresponding abbreviated and full forms as very closely related, unless these are the only differences in an otherwise similar, longer segment. Some abbreviations can also be seen as synonyms, for example "phone" for "telephone", especially where they have come into very general usage. Character-wise, this can be seen as an omission (see section 5.2.), but hides the fact that there is a very strong semantic relationship between such terms—they are indeed interchangeable as synonyms.

Related to the use of synonymy is segments that are mutual **paraphrases**. Cases with **active/passive** paraphrases, or paraphrases only due to changes in **word order** were classified separately. Although English does not have a particularly free word order, certain differences in word orders are possible. Post-positive adjectives were noted in this category, for example.

These categories therefore represent cases where the seemingly different source texts are semantically equivalent or very similar.

5.2. Omission

Another large category identified is where one or more omissions are the most notable difference between the two seemingly unrelated source texts. Some examples are shown in table 4.

In some cases two source texts differed in that one used a pronoun instead of a noun phrase. In most cases, however, the reason for the omission was not obvious when viewed out of context in this fashion.

5.3. Text normalisation

In the test corpora, text normalisation issues relating to letter case and character representation comprised a small number of cases. Some examples are shown in table 5.

Case folding is often used in information retrieval systems to improve recall and reduce index size. (Manning et al., 2009) In the corpora investigated in this study, letter **case** was the main reason for low source text similarity in only a very small number of cases.

The lack of **normalisation** of character representation was identified as the primary reason in a few cases in the GNOME corpus, for example the use of the single-character ellipsis (Unicode value 2026, HORIZONTAL ELLIPSIS) vs. three separate full stop characters. While the Hungarian target texts in the GNOME corpus consistently uses only single-character ellipses, the English text also uses three full-stop characters in some segments, which causes the mismatch in the source text, especially for shorter segments.

In the GNOME corpus, the underscore character "_" is used to indicate that the following character is a mnemonic (e.g. "_File" is presented in the GUI as "File" to indicate that `Alt+F` activates the menu). Some variation in the selection of mnemonic characters is expected for software GUIs. This only affects at most two positions in the source string for the comparison as done in this study (one insertion point and one deletion point), and therefore only resulted in missed TM opportunities in strings with five or less characters. (An edit distance of 2 in a string of length 5 results in a similarity of 60%—less than our threshold.) These, along with similar issues such as the use of straight vs. curly quotation marks, often contributed as secondary reasons for missed opportu-

Source/Target text	Source/Target text	Similarity	Category
blank vide	empty vide	0% 100%	synonym
Type a new shortcut Saisissez un nouveau raccourci clavier	Type a new accelerator Saisissez un nouveau raccourci clavier	55% 100%	synonym
No such book **C**e carnet d'adresses n'existe pas	Address book does not exist **L**e carnet d'adresses n'existe pas	19% 96%	paraphrase
UPS Onduleur	Uninterruptible power supply Onduleur	3.6% 100%	abbreviation
The product is sold fresh or chilled.' Csak friss és hűtött termék árusítható."	Only fresh or chilled product may be sold.' Csak friss és hűtött termék árusítható."	26% 100%	reordering

Table 3: Examples of semantic categories

Source/Target text	Source/Target text	Similarity	Category
Name Nom	Last **Name** Nom	44% 100%	omission
Discard Annuler les modifications	Discard **Changes** Annuler les modifications	47% 100%	omission
Having any of the following: Rendelkez**i**k az alábbi jellemzők bármelyikével:	Either of the following **characteristics**: Rendelkez**ne**k az alábbi jellemzők bármelyikével:	38% 95%	omission
It shall apply from 1 January 2004. Ezt a határozatot 200**4**. január 1-jétől kell alkalmazni.	**This Decision** shall apply until 1 January 2005. Ezt a határozatot 200**5**. január 1-jétől kell alkalmazni.	57% 98%	omission

Table 4: Examples of omission

nities. In combination with other, more substantial effects on source text similarity, the similarity score for two source strings are more likely to drop below the threshold.
The mismatch of XML entities with the characters they represent (e.g. " for ") cause a larger variance in string length (6 characters vs. 1), which is more likely to lower the source string similarity to below a threshold.

5.4. Misalignment

A large category exclusive to the DGT corpus is where the source-target segments are misaligned. Some examples are shown in table 6. This was more frequent than expected, since the authors mentioned that "[...] the alignment quality was found to be very good, with only few errors." (Steinberger et al., 2012). Different country names were very frequently extracted (in both language pairs) with identical translations. This suggests mass-misalignment of some lists of country names. However, since we employed no linguistic knowledge of the target languages, the category for misalignment was reserved only for the most obviously misaligned segments—mostly identified by very large differences in segment length.
No alignment errors were expected in the GNOME corpus, and none were found.

6. Discussion

In this section, we briefly consider some possible directions for future work that might improve TM systems for the categories mentioned above.

The number of misalignments in the DGT corpus was unexpected. It would be unrealistic to expect a TM system to be able to retrieve such suggestions. Even though the target texts might be useful by accident in a particular case, no suggestion is made to improve this inside a TM system. A proper investment in sentence alignment and verification of the alignment would be the better place to focus attention to solve this problem.

Semantic similarity measures might be of assistance to improve cases where synonyms, abbreviations and paraphrases occur. Synonym handling might be improved with the use of thesauri, or by synonym extraction using the TM itself (van der Plas and Tiedemann, 2006). Matters of word order might be solved by using methods not as strongly tied to the order of characters/words. This was already suggested in Baldwin (2009), but more investigation will be required, as this category was not very big.

The omissions category suggests using three operation edit distance rather than four operation edit distance, as attempted in Baldwin (2009).

The size of the text normalisation category suggests that efforts in this regard will not improve recall substantially. However, since it played a secondary role in missing some opportunities, it should not be disregarded completely. Unicode normalisation forms (Davis et al., 2009) might provide solutions to several solutions of character representation. At least some normalisation issues are already handled by

Source/Target text	Source/Target text	Similarity	Category
Because "{1}" À cause de «{1} ».	Because "{1}". À cause de «{1} ».	50% 100%	representation
TOTAL ÖSSZESEN	Total ÖSSZESEN	20% 100%	letter case

Table 5: Examples of text normalisation issues

Source/Target text	Source/Target text	Similarity	Category
Read: helyesen:	the name of the first independent customer in the Community to which the invoice is issued directly by the sales company; helyesen:	4% 100%	misalignment
Egypt Égypte	Ecuador 69,6 Egypt Égypte	28% 100%	misalignment
Lebanon Liban	Lesotho Liban	29% 100%	*misalignment (not counted)

Table 6: Examples of misalignment

translation tools in use.[2]

7. Conclusion

This study set out to identify reasons why existing TM systems fail to provide certain useful suggestions in a TM database. There are several reasons, not all of them equally frequent in the test data.

The study investigated results across two linguistically unrelated languages in two data sets from different domains with very different properties. Although there is some variation between these four combinations, it was clear that a similarity metric based on edit distance is likely to miss several useful suggestions. The largest category of missing suggestions were those with high semantic similarity between the source texts. Omission was another large category across all the data sets.

Returning to our research question, we showed that improvements in recall are still possible, and we proposed a categorisation that would facilitate a further, deeper analysis and provided suggestions towards finding solutions. With a similarity metric based on edit distance, an easy way to improve recall is to lower the fuzzy match threshold. This simply trades precision for recall. An attempt to improve a similarity method would need to be tested on a full corpus—not only these previously missed opportunities—especially the effect on precision.

8. Acknowledgement

This research was supported in part by funding from the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight) and the Academy of African Languages and Science Strategic Project of the University of South Africa.

[2]e.g. WordFast can ignore case differences when matching, while still affecting ranking (Moslem, 2012)

9. References

Andre, R. (2013). Wordfast Anywhere FAQ. http://www.wordfast.net/wiki/Wordfast_Anywhere_FAQ. Accessed: 2014-03-19.

Baldwin, T. (2009). The hare and the tortoise: speed and accuracy in translation retrieval. *Machine Translation*, 23:195–240.

Davis, M., Whistler, K., and Dürst, M. (2009). Unicode normalization forms. Technical Report Unicode Standard Annex 15, Unicode Consortium.

Lagoudaki, E. (2006). Translation memories survey 2006: Users' perceptions around TM use. In *Proceedings of the ASLIB International Conference Translating & the Computer*, volume 28.

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

Macklovitch, E. and Russell, G. (2000). What's been forgotten in translation memory. *Envisioning Machine Translation in the Information Future*, pages 205–207.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Morado Vázquez, L. (2012). *An empirical study on the influence of translation suggestions' provenance metadata*. Ph.D. thesis, University of Limerick.

Moslem, Y. (2012). Applying penalties in WFP. http://www.wordfast.net/wiki/Applying_Penalties_in_WFP. Accessed: 2014-03-19.

O'Brien, S. (2007). Eye-tracking and translation memory matches. *Perspectives: Studies in translatology*, 14(3):185–205.

SDL plc. (2013). About translation memory matches. http://producthelp.sdl.com/SDL%20Trados%20Studio/client_en/Edit_View/TMs/EVWorkingwithTMsAbout_

Translation_Memory_Matches.htm. Accessed: 2014-03-19.

Somers, H. (2003). *Computers and translation: a translator's guide*, volume 35. John Benjamins Publishing Company.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2012). DGT-TM: A freely available translation memory in 22 languages. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

van der Plas, L. and Tiedemann, J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 866–873, Stroudsburg, PA, USA. Association for Computational Linguistics.