Towards New Generation Translation Memory Systems

Nikola Spasovski, Ruslan Mitkov

Research group in Computational Linguistics
University of Wolverhampton
{n.spasovski,r.mitkov}@wlv.ac.uk

Abstract

Despite the enormous popularity of Translation Memory systems and the active research in the field, their language processing features still suffer from certain limitations. While many recent papers focus on semantic matching capabilities of TMs, this planned study will address how these tools perform when dealing with longer segments and whether this could be a cause of lower match scores. An experiment will be carried out on corpora from two different (repetitive) domains. Following the results, recommendations for future developments of new TMs will be made.

1 Introduction

Translation Memory systems are one of the many translation technology tools available in the translation market. Their main purpose is to accelerate translators' productivity and improve the overall translation process. As many authors recognise, Translation Memories have achieved remarkable success among both translation companies and individual translators, and have positively impacted translators' work. (Lagoudaki (2008); Bowker (2008)). Their popularity is evidenced by the results of the 2018 Language Industry Survey[1] – Expectations and Concerns of the European Language Industry, which show that the use of computer assisted translation (CAT) tools is widespread in both language service companies and in individual professionals/freelancers, with only less than 1% of companies reporting that they do not use CAT tools, and only around 13% of individual language professionals.

By virtue of the lively interest CAT tools have received from the industry and academia over the last three decades, the software and technical features of these tools have also been evolving at rapid pace, thus successfully 'surpassing' rival tools and technologies in the extremely competitive translation market. Furthermore, as a consequence of an increasing demand for this type of technology, we have witnessed a diversification and multiplication of the range of TM systems on the market today.

The concept of Translation Memory is a simple one. It can be defined as a database used to store previously translated segments. This database consists of parallel texts and their translations, which are segmented and then aligned according to a sentence-based method. Whenever there is an equal or similar phrase to be translated, these segments are offered as exact or fuzzy matches. The chances of a retrieval increase as the translator stores more translations in the database. Nonetheless, if there are no exact/fuzzy matches to be suggested, nowadays the TM can offer 'similar matches', or a match from the Machine Translation extension of most CAT tools. It is left to the translator to decide whether the 'similar' or MT matches are useful, and these can be used in the translation by editing the differences.

The sentence matching in TMs is based on 'string-edit' distance, or more precisely, the 'Levenshtein distance' (Levenshtein, 1966). Levenshtein distance, in the words of Somer (2003), is the 'minimum number of changes (e.g. insertions, deletions and substitutions) needed to replace one segment with another'. In TM terms, this means that the database segments/chunks are compared to the sentence to be translated based on the number of changed characters, or distance, between words.

Using 'string-edit' distance makes TMs practical for repetitive texts, such as technical ones, that abound in similar terminology as these systems simplify the task of reusing previously translated content. In line with the aforementioned, Zaretskaya (2016) notes that translators of

[1] The European Language Industry Survey can be accessed on the following link.

technical content are the most likely candidates to benefit from TM tools, followed by legal translators and those translating financial and/or marketing documents.

Nevertheless, many researchers (Macklovitch and Russell (2000); Pekar and Mitkov (2007); Reinke (2013); Baquero and Mitkov (2017)), Mitkov et al. (2021) point out the main limitation of the Levenshtein distance - the lack of language processing functionalities - which stands in the way of TMs recognising matches between syntactically different but semantically equal segments.

Consequently, research including TM systems has been constantly sparked by a single question: *How to improve these systems so that translators can benefit from them even more?*

2 Related work

In academia, TMs have been approached from both the technical (NLP) aspect, which includes work on their capabilities of semantic matching, and the user aspect, which addresses users' necessities regarding the tools.

From the user aspect, Lagoudaki (2008) performed extensive research on users' needs regarding translation memories. The results of her work showed that translators would not accept any changes in TM tools if the existing processing speed were affected and if those changes resulted in additional time–consuming tasks in the translation process. Furthermore, translators consider that changes are not necessary unless they make the software perform tasks faster and better than a human. However, most experienced respondents answered that 'the system's computational efforts must stop at the point where the imagination of the translator risks being compromised'. Nevertheless, young translators believe that the matching capabilities of a TM system could be improved with the implementation of machine translation techniques.

Zarteskaya et al. (2016) conducted a user survey on CAT tools, which included a section dedicated to the 'functionality' features of the tools. According to the results obtained by the survey, speed is a fundamental part of translators' work. Hence, the most important characteristics that a TM system should have are a high working speed, followed by a user-friendly interface, and ease of use.

On the NLP side, Pekar and Mitkov (2007) proposed a new generation of translation memory capable of performing semantic matching of sentences. Several studies that consisted of paraphrase recognition experiments on TMs were conducted as an answer to this semantic matching approach. Among them are: Marsye (2011), Timonera and Mitkov (2015), Chatzitheodorou (2015).

In Marsye (2011), authors assessed how TM systems would perform in terms of suggesting possible match segments that were paraphrases of the source segment. In order to identify synonym words and recognise paraphrases, the author suggests using WordNet as an extension of the TMs as a possible solution.

Later work included Chatzitheodorou (2015), who used the NooJ module to create equivalent paraphrases from the source texts of a TM. He proved that this method could increase fuzzy matches; however, the implemented module failed to paraphrase certain chunks.

Furthermore, in order to improve match retrieval, Timonera and Mitkov (2015) carried out experiments with clause splitting as well as paraphrasing. Following the experiments, the authors came to the conclusion that paraphrasing and the use of a Paraphrase database (PPDB) does result in an improved match retrieval.

More recent attempts to improve translation memory systems involve experiments with neural networks and deep learning methods (Mitkov 2021). Among them, Ranasinghe, et al. (2020) are worth mentioning, who employed NLP and DL techniques such as word and sentence embeddings to match segments from the TM database instead of the previously used Levenshtein (edit) distance.

Although the aforementioned studies performed reasonably well and enhanced the capabilities of semantic matching, all of them present limitations. Most have not been tested in real-life scenarios, and some even lack information on the time required by the system to suggest a segment from the TM database. Considering all user-conducted surveys pointed out speed as the most valued component of the tools, it remains to be seen whether the - updated - system outperforms Translation Memories in terms of speed and is able to offer matches faster.

Moreover, two of the experiments see storage space-related issues. In one of the studies, as the database grows, the TM processing/retrieval time becomes slower, while the second study found that the larger the database, the more RAM space is needed. Thus, applied to a real-life scenario, translators' speed would be affected and additional storage space would be required.

3 Methodology

Across the different literature reviewed, at the time of writing this research proposal, it was found that there exists little to no work that explores in depth the topic of translation memories and their performance on longer segments. This fact stands out when we consider that some of the available tools, such as MateCat and MemoQ, fail to return a high-percent fuzzy match if the segment is significantly longer. Hence, a methodology for carrying out the study is designed:

A known deficiency, an analysis of performance in a defined setting (long repetitive segments in financial and scientific texts), and possible suggestions for what to come afterwards.

The overall goal of this project is to evaluate to what degree TMs fail in retrieving matches for longer (repetitive) segments by using their in-built text processing algorithms, hence; the main research question in this study will be *how can matching (of longer segments) in existing translation memories be improved, making the tools more useful for translators?*

3.1 Language data (Parallel Corpora as Translation Memories)

The working languages of the study are English and Spanish – each one will be treated as a source language. To this end, available parallel corpora from two different domains will be used. The corpora will be imported and aligned. They will then be processed into the database as a TM, so that they can be properly exploited. The first corpus is the European Central Bank corpus, which contains financial vocabulary extracted from the Bank's website, and the second corpus is going to be selected in a later stage of the study. It is expected that the focus will be on a medical domain (anatomy, more specifically). The main assumption is that certain domains are more repetitive, and this characteristic influences the final TM output in terms of retrieving higher fuzzy matches. Therefore, as part of the corpora preprocessing, we will iterate over them and measure the degree of segment repetition. The repetition measurement will be performed using n-grams (trigrams, fourgrams) in the Natural Language Toolkit (NLTK) library in Python. After performing the repetition measurement, shorter and longer segments will be divided and experimented with. For the purposes of this research, the reference size will be 1700/1800 characters, so that a "short segment" will be shorter than 1700 and a "longer one" will be equal or longer than 1700. Given that existing corpora often present some issues, another procedure as part of the text-preprocessing will be to perform data cleaning. This will include three steps:
1. check whether both corpora in Spanish and English use the same XML tags;
2. check for mismatches in date format, numbers (whether they are written with digits or letters);
3. check for empty segments in one of the languages and/or segments in different languages than the required.

These details are considered important since they could influence the fuzzy match percentage of the TMs to be used. For the data cleaning, one of the publicly available sources suggested by Barbu (2017) will be used.

3.2 Experiments

In order to answer the research question, this study will assess the TMs' output and try to find ways to improve them. One of the most important components of the investigation to focus on will be the fuzzy matches. We will seek to investigate and compare the attributes and translation output of several TM systems: commercial (e.g. MemoQ, Wordfast), open (e.g. SmartCAT, MateCat), as well as TM tools with web interface (e.g. Memsource, Matecat). Both word-based and character-based Levenshtein distance typically employed in commercial TM systems will be put to the test. Ideally, in order to assess performance, the matching threshold will be set between 70 and 80 percent. We will observe how the fuzzy match algorithm works with longer segments and whether a greater length could be a reason for a lower score match. If this is the case, we will measure to what degree TMs fail in retrieving matches for longer segments. Additionally, a comparison regarding speed and matching accuracy between both types of segments will be carried out.

Finally, the influence of formatting and tags on the match retrieval and how they could impact the semantic recognition when they are not well placed within long (or short) segments will be analyzed.

3.3 Future work

Following these experiments, the typical errors will be categorised and recommendations for the development of a new generation of TM systems (that integrate linguistic knowledge) will be made. We will be aiming for this work to be a base for NLP engineers seeking to improve TMs, as well as for other researchers in translation technologies working with longer or repetitive segments (from finance and science) in general and to (eventually) prepare a basis for future interpreting memory tools or other language tools that rely on TMs.

References

2018 Language Industry Survey – Expectations and Concerns of the European Language Industry

Baquero Silvestre A., Mitkov R. (2017). Translation Memory Systems have a long way to go. In proceedings of the Workshop Human-Informed Translation and Interpreting Technology https://doi.org/10.26615/978-954-452-042-7_006

Barbu E. (2017). Ensembles of Classifiers for Cleaning Web Parallel Corpora and Translation Memories. In proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP) https://doi.org/10.26615/978-954-452-049-6_011

Bowker L. (2008) Computer - Aided Translation Technology: A Practical Introduction. Ottawa: University of Ottawa Press

Chatzitheodorou K. (2015). Improving translation memory fuzzy matching by paraphrasing. In proceedings of the Workshop on Natural Language Processing for Translation Memories (NLP4TM), pages 24–30.

European Central Bank corpus J. Tiedemann, (2012), Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)

Lagoudaki E. (2008). Expanding the Possibilities of Translation Memory Systems From the Translator's Wishlist to the Developer's Design. PhD thesis, Imperial College London

Levenshtein, Vladimir I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In Soviet physics doklady , volume 10, 707–710.

Macklovitch E, Russell G. (2000). What’s been Forgotten in Translation Memory. In: Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future, London, UK, pages 137–146

Marsye A. (2011). Towards a New Generation Translation Memory: A Paraphrase Recognition Study for Translation Memory System Development. Master thesis, University of Wolverhampton

Mitkov, R. (2021). Translation Memory. In S. Deane-Cox and A. Spiessens (Eds), The Routledge Handbook of Translation and Memory. Basingstoke: Routledge.

Pekar V, Mitkov R. (2007). New Generation Translation Memory: Content - Sensitive Matching. In proceedings of the 40th anniversary congress of the Swiss association of translators, terminologists and interpreters.

Ranasinghe T, Mitkov R, Orăsan C and Caro Quintana R. (2020). Semantic Textual Similarity based on Deep Learning: Can it improve matching and retrieval for Translation Memory tools?

Reinke U. (2013). State of the Art in Translation Memory Technology. In proceedings of the Workshop on Natural Language Processing for Translation Memories (NLP4TM), pages 17–23

Somers H. (2003). Computers and Translation: A Translator’s Guide. Amsterdam and Philadelphia: John Benjamins.

Timonera K, Mitkov R. (2015). Improving Translation Memory Matching through Clause Splitting. In proceedings of the Workshop on Natural Language Processing for Translation Memories (NLP4TM), pages 17–23

Zaretskaya A, Corpas G, Seghiri M. (2016). User Perspective on Translation Tools: Findings of a User Survey

Zaretskaya A. (2016). Translators’ requirements for translation technologies: user study on translation tools. PhD thesis, Universidad de Málaga