Improving translation memory fuzzy matching by paraphrasing

Konstantinos Chatzitheodorou
School of Italian Language and Literature
Aristotle University of Thessaloniki
University Campus
54124 Thessaloniki, Greece
chatzik@itl.auth.gr

Abstract

Computer-assisted translation (CAT) tools have become the major language technology to support and facilitate the translation process. Those kind of programs store previously translated source texts and their equivalent target texts in a database and retrieve related segments during the translation of new texts. However, most of them are based on string or word edit distance, not allowing retrieving of matches that are similar. In this paper we present an innovative approach to match sentences having different words but the same meaning. We use NooJ to create paraphrases of Support Verb Constructions (SVC) of all source translation units to expand the fuzzy matching capabilities when searching in the translation memory (TM). Our first results for the EN-IT language pair show consistent and significant improvements in matching over state-of-the-art CAT systems, across different text domains.

1 Introduction

The demand of professional translation services has been increased over the last few years and it is forecast to continue to grow for the foreseeable future. Researchers, to support this increasing, have been proposed and implemented new computer-based tools and methodologies that assist the translation process. The idea behind the computer-assisted software is that a translator should benefit as much as possible from reusing translations that have been human translated in the past. The first thoughts can be traced back to the 1960s when the European Coal and Steel Community proposed the development of a memory system that retrieves terms and their equivalent contexts from earlier translations stored in its memory by the sentences whose lexical items are close to the lexical items of the sentence to be translated (Kay, 1980).

Since then, TM systems have become indispensable tools for professional translators who work mostly with content that is highly repetitive such as technical documentation, games and software localization etc. TM systems typically exploit not only exact matches between segments from the document to be translated with segments from previous translations, but also approximate matches (often referred to as fuzzy matches) (Biçici and Dymetman, 2008). As concept, this technique might be more useful for a translator because all the previous human translations become a starting point of the new translation. Furthermore, the whole process is speeded up and the translation quality is more consistent and efficient.

The fuzzy match level refers to all the necessary corrections made by a professional translation in order to make the retrieved suggestion to meet all the standards of the translation process. This effort is typically less than translating the sentence from scratch. To help the translator, CAT tools suggest or highlight all the differences or similarities between the sentences, penaltizing as well the match percent in some cases. However, given the perplexity of a natural language, for similar, but not identical sentences the fuzzy matching level sometimes is too low and therefore the translator is confused.

This paper presents a framework that improves the fuzzy match of similar, but not identical sentences. The idea behind this model is that Y2 which is the translation of Y1 can be the equivalent of X1 given that X1 has the same meaning

Proceedings of the Workshop on Natural Language Processing for Translation Memories (NLP4TM), pages 24–30,
Hissar, Bulgaria, Sept 2015.

with Y1. We use NooJ to create equivalent paraphrases of the source texts to improve as much as possible the translation fuzzy match level given that they share the same meaning but not the same lexical items. In addition to this, we investigate the following questions: (1) is the productivity of the translators improved? (2) are SVC widespread to merit the effort to tackle them? These questions are answered using human centralized evaluations.

The rest of the paper is organized as follows: Section 2 discusses the past related work, section 3 the theoretical background, section 4 the conceptual background as well the architecture of the framework. Section 5 details the experimental results and section 5 the plans for further work.

2 Related work

There has been some work to improve the translation memory matching and retrieval of translation units when working with CAT tools (Koehn and Senellart, 2010; He at al, 2010a; Zhechev and van Genabith, 2010; Wang et al., 2013). Such works aim to improve the machine translation (MT) confidence measures to better predict the human effort in order to obtain a quality estimation that has the potential to replace the fuzzy match score in the TM. In addition to this, these techniques have an effect only in improvement of the MT raw output and not in improvement of fuzzy matching.

A common methodology that gives priority to the human translations is to search first for matches in the project TM. When no such close match is found in the TM, the sentence is machine-translated (He at al, 2010a; 2010b). In a somewhat similar spirit, other hybrid methodologies combine techniques at a sub-sentential level. Most of them, use as much as possible human translations for a given sentence and the unmatched lexical items are machine translated in the target language using a MT system (Smith and Clark, 2009; Koehn and Senellart, 2010; He at al, 2010a; Zhechev and van Genabith, 2010; Wang et al., 2013). Towards the improving of the quality of the MT output, researchers have been using different MT approaches (statistical, rule-based or example-based) trained either on generic or in-domain corpora. Another innovative idea has been proposed by Dong et al. (2014). In their work, they use a lattice representation of possible translations in a monolingual target language corpus to find the potential candidate translations.

On the other hand, various researchers have focused on semantics or syntactic techniques towards improving the fuzzy matching scores in TM but the evaluations they performed were shallow and most of the time limited to subjective evaluation by authors. Thus, this makes it hard to judge how much a semantically informed TM matching system can benefit a professional translator. Planas and Furuse (1999) propose approaches that use lemma and parts of speech along with surface form comparison. In addition to this syntactic annotation, Hodász and Pohl (2005) also include noun phrase (NP) detection (automatic or human) and alignment in the matching process. Pekar and Mitkov (2007) presented an approach based on syntax driven syntactic analysis. Their result is a generalized form after syntactic, lexico-syntactic and lexical generalization.

Another interested approach, similar to ours, has been proposed by Gupta and Orasan (2014). In their work, they generate additional segments based on the paraphrases in a database while matching. Their approach is based on greedy approximation and dynamic programming given that a particular phrase can be paraphrased in several ways and there can be several possible phrases in a segment which can be paraphrased. It is an innovative technique, however, paraphrasing lexical or phrasal units in not always safe and in some cases, it can confuse rather than help the translator. In addition to this, a paraphrase database is required for each language.

Even if the experimental results show significant improvements in terms of quality and productivity, the hypotheses are produced by a machine using unsupervised methods and therefore the post-editing effort might be higher comparing to human translation hypotheses. To the best of our knowledge, there is no similar work in literature because our approach does not use any MT techniques given that target side of the TM remains “as is”. To improve the fuzzy matching, we paraphrase the source translation units of the TM, so that a higher fuzzy match will be identified for sentences sharing the same meaning. Therefore, the professional translator is given a human translated segment that is the paraphrase of the sentence to be translated. This ensures that no out-of-domain lexical items or no machine translation errors will appear in the hypotheses, making the post-editing process trivial.

3 Theoretical background

There are several implementations of the fuzzy match estimation during the translation process, and commercial products typically do not disclose the exact algorithm they use (Koehn, 2010). However, most of them are based on the word and/or character edit distance (Levenshtein distance) (Levenshtein, 1966) i.e., the total number of deletions, insertions, and substitutions in order the two sentences become identical (Hirschberg, 1997).

For instance, the word-based string edit distance between sentence (1) and (2) is 70% (1 substitution and 3 deletions for 13 words), and the character-based string edit distance is 76% (14 deletions for 60 characters) without counting whitespaces based on Koehn's (2010) formula for fuzzy matching. This is a low score and many translators may decide not to use it and therefore not to gain from it.

(1) Press ' Cancel ' to **make the cancellation** of your personal information .

(2) Press ' Cancel ' to **cancel** your personal information .
(3) Premere ' Cancel ' per cancellare i propri dati personali .

(4) Press ' Cancel ' to **cancel** your booking information .

In this case, according to methodologies proposed by researchers of this field, this sentence will be sent for machine translation given the low fuzzy match score and then it should be post-edited. Otherwise, the translator should translate it from scratch. However, this is not always safe, given that in many cases post-editing MT output requires more time than translating from scratch.

Observing the differences between sentences (1) and (2) one can easily conclude that they share the same meaning although they don't share the same lexical items. This happens because of their syntax. In more detail, sentence (1) contains a SVC while sentence (2) contains its nominalization. An EN-IT professional translator can benefit from our approach by accepting the sentence (3) as the equivalent translation of the sentence (1).

SVCs, like *make a cancellation*, are verb-noun complexes which occur in many languages. Form a syntactic and semantic point of view they act in the same way as multi-word units. Their meaning is mainly reflected by the predicate noun, while the support verb is often semantically reduced. The support verb contributes little content to its sentence; the main meaning resides with the predicate noun (Barreiro, 2008).

SVCs include common verbs like *give*, *have*, *make*, *take*, etc. Those types of complexes can be paraphrased with a full verb, maintaining the same meaning. While support verbs are similar to auxiliary verbs regarding their meaning contribution to the clauses in which they appear, support verbs fail the diagnostics that identify auxiliary verbs and are therefore distinct from auxiliaries (Butt, 2003).

SVCs challenge theories of compositionality because the lexical items that form such constructions do not together qualify as constituents, although the word combinations do qualify as catenae. The distinction of a SVC from other complex predicates or arbitrary verb-noun combinations is not an easy task, especially because their syntax that is not always fixed. Except of some cases, they appear with direct object (e.g. *to make attention*) or with direct object (e.g. *to make a reservation*) (Athayde, 2001).

Our approach paraphrases SVCs found in the source translation units of a TM in order to increase the fuzzy matching between sentences having the same meaning. It is a safe technique because the whole process has no effect on the target side of the TM translation units. Hence, the translators benefit only from human translation hypotheses that usually are linguistically correct.

In our example, an EN-IT translator will receive an exact match during his performance when translating the sentence (1) given the English sentence (2) and its Italian equivalent (sentence (3)) that is included in the TM. In addition to this, in case of translating the sentence (4), the fuzzy match score would be around 90% (1 substitution for 10 words) comparing to 61% with no-paraphrase (2 substitution and 3 deletions for 13 words). Other than fuzzy match, according to Barreiro (2008) machine-translation of SVCs is hard, so the expected output from the machine will not be good enough. In our example, "cancel" can be either a verb or noun.

4 Conceptual background

As already discussed, paraphrasing a SVC can increase the fuzzy match level during the translation process. This section details the pipeline of modules towards the paraphrase of the TM source translation units.

4.1 NooJ

The main component of our framework is NooJ (Silberztein, 2003). NooJ is a linguistic development environment used to construct large-coverage formalized descriptions of natural languages, and apply them to large corpora, in real time. The

module consists of a very large lexicon, along with a large set of local grammars to recognize named entities as well as unknown words, word sequences etc. These resources have been obtained from OpenLogos, an old open source rule-based MT system (Scott and Barreiro, 2009). In NooJ, an electronic dictionary contains the lemmas with a set of information such as the category/part-of-speech (e.g. V for verbs, A for adjectives etc.), one or more inflectional and/or derivational paradigms (e.g. how to conjugate verbs, how to lemmatize or nominalize them etc.), one or more syntactic properties (e.g. +transitiv for transitive verbs or +PREPin etc.), one or more semantic properties (e.g. distributional classes such as +Human, domain classes such as +Politics) and finally, one or more equivalent translations (+IT="translation equivalent"). Figure 1 illustrates typical dictionary entries.

```
artist,N+FLX=TABLE+Hum
cousin,N+FLX=TABLE+Hum
pen,N+FLX=TABLE+Conc
table,N+FLX=TABLE+Conc
man,N+FLX=MAN+Hum
```

Figure 1: Dictionary entries in NooJ for nouns.

4.2 Paraphrasing the source translation units

The generation of the TM that contains the paraphrased translation units is straightforward. The architecture of the process which is summarized in Figure 2, is performed in three pipelines:

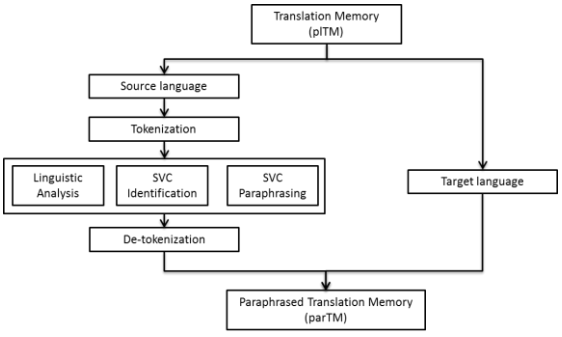

Figure 2: Pipeline of the paraphrase framework.

The first pipeline includes the extraction of the source translation units of a given TM. The target translation units are protected so that they will not be parsed by the framework. This step also includes the tokenization process. Tokenization of the English data is done using Berkeley Tokenizer (Petrov et al., 2006). The same tool is also used for the de-tokenization process in the last step.

Then, all the source translation units pass through NooJ to identify the SVCs using the local grammar of Figure 3. To do so, NooJ first pre-processes and analyses the text based on specific dictionaries and grammars attached in the module. This is a crucial step because if the text is not correctly analyzed, the local grammar will not identify all the potential SVCs and therefore there will not be any gain in terms of fuzzy matching. Once the text is analyzed, all the possible SVCs are identified and hence paraphrased.

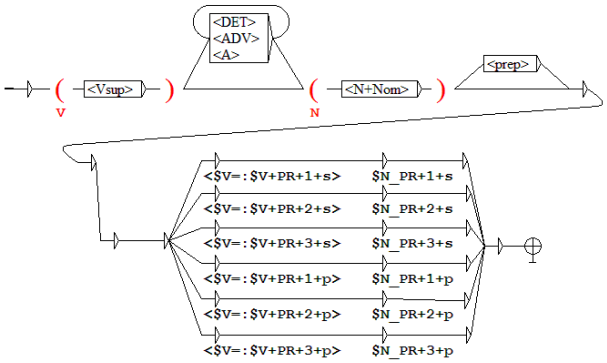

Figure 3: Local grammar for identification and paraphrasing of SVCs.

In more detail, the local grammar checks for a support verb followed by a determiner, adjective or adverb (optionally), a nominalization and optionally by a preposition, and generates the verbal paraphrases in the same tense and person as the source. We should notice that this graph recognizes and paraphrases only SVCs in simple present indicative tense. However, our NooJ module contains grammars created for the all the other grammatical tenses and moods that follow the same structure. The elements in red colors characterize the variables as verb and predicate nouns. The elements **<$V=:$V+PR+1+s>**, and **$N_PR+1+s** represent lexical constraints that are displayed in the output, such as specification of the support verb that belongs to a specific SVC. These particular elements refer to the first person singular of the simple present tense. The predicate noun is identified, mapped to its deriver and displayed as a full verb while the other elements of the sentence are eliminated. The final output of NooJ is a sentence that contains the paraphrase instead of the SVCs, were applicable.

The last pipeline contains the de-tokenization as well as the concatenation of the paraphrased

translation units in the original TM, if any. The paraphrased translation units have the same proprietaries, tags etc., as the original units.

This TM should be imported and used in the same way as before in all CAT tools. As of now, our approach can be applied only to TMs that have the English language as source. As mentioned earlier, there is no limit for the target language given that we apply our approach only to the source language translation units.

5 Experimental results

The aim of this research is to provide translators with fuzzy match scores higher than before in case the TM contains a translation unit which has the same meaning with the sentence to be translated. Given that there is no automatic evaluation for this purpose, we formulate this as a ranking problem. In this work, we analyze a set of 100 sentences from automotive domain and 100 from IT domain to measure the difference of the fuzzy match scores between our approach (parTM) and the conversional translation process, where a plain TM is used (plTM). This test set, was selected manually in order to contain SVCs in order to ensure that each sentence contains at least one SVC.

Our method has been applied to a TM which contained 1025 EN-IT translation units. Our module recognized 587 SVCs, so the generated TM (parTM) was contained 1612 translation units (1025 original + 587 paraphrases). The TM contains translations that have been taken from a larger TM based on the degree of fuzzy match that at least meets the minimum threshold of 50%. To create the analysis report logs we used Trados Studio 2014[1].

The results of both analyses are given in Table 1.

Our paraphrased TM attains state-of-the-art performance on increasing the fuzzy match leveraging. It is interesting to note that the highest gains are achieved in the low fuzzy categories (0%-74%). However, we achieve extremely high numbers in other categories. Our approach improves the scores by 17% in 100% match category, 5% in category 95% - 99%, 6% in category 85% - 94%, 28% in category 75% - 84% and finally, 27% in category 0%-74% (No match + 50%-74%). This is a clear indication that paraphrasing of SVCs significantly improves the retrieval results and hence the productivity.

Fuzzy match category	**plTM**	**parTM**
100%	14	48
95% - 99%	23	32
85% - 94%	18	29
75% - 84%	51	38
50% - 74%	32	18
No Match	62	35
Total	**200**	**200**

Table 1: Statistics for experimental data

To check the quality of the retrieved segments human judgment was carried by professional translators. The test set consist of retrieved segments with fuzzy match score >=85% (108 segments). The motivation for this evaluation is twofold. Firstly to show how much impact paraphrasing of SVCs has in terms of retrieval and secondly to see the translation quality of those segments which the fuzzy match score is improved because of the paraphrasing process.

According to translators, paraphrasing helps and speeds up the translation process. Moreover, the fact that the target segments remain "as is" encourage them to use it without a second thought.

Figure 4 shows two cases where translators selected to use segments from the parTM. We can see that paraphrasing not only helps to increase the retrieving but also ensures that the proposed translation is a human translation, so no errors will appear and less post editing is required in case of not equal to 100%.

While there are some drops in terms of fuzzy match improvement, our system presents few weaknesses. Most of them regard the out-of-vocabulary words during the analysis process by NooJ. Although our NooJ module contains a very large lexicon, along with a very large set of local grammars to recognize and paraphrase SVCs, a few translation units (6 segments) were not paraphrased. In addition to this, 2 segments were paraphrased incorrectly. This happens because they contain either out-of-vocabulary words or due to their syntax complexity. This is one of our approach's weaknesses that will be addressed for future projects.

Seg:	**Make sure** that the brake hose is not twisted.
TMsl:	**Ensure** that the brake hose is not twisted
TMtg	**Assicurarsi** che il tubo flessibile freni non sia attorcigliato.

[1] http://www.sdl.com/cxc/language/translation-productivity/trados-studio/

parTMsl:	**Make sure** that the brake hose is not twisted.

Seg:	CAUTION: You must **make the istallation** of the version 6 of the software.
TMsl:	CAUTION: You **must install** the version 6 of the software.
TMtg	ATTENZIONE: Si **deve installare** la versione 6 del software.
parTMsl	CAUTION: You must **make the istallation** of the version 6 of the software.

Figure 4: Accepted translations.

6 Conclusion

In this paper, we have presented a method that improves the fuzzy match of similar, but not identical sentences. We use NooJ to create equivalent paraphrases of the source texts to improve as much as possible the translation fuzzy match level given that the meaning is the same but they don't share the same lexical items.

The hybridization strategy implemented has already been evaluated with different experiments, translators, text types and language pairs, which showed that it is very effective. The results show that for all fuzzy-match ranges our approach performs markedly better than the plain TM for different fuzzy match levels, especially for low fuzzy match categories. In addition to this, the translators' satisfaction and trust is abundant comparing to MT approaches.

In the future, we will continue to explore ways paraphrasing of other support verbs and other support languages as well. Last but not least, a paraphrase framework to the target sentence may improve even more the quality of translations.

References

Athayde M. F. 2001. *Construções com verbo-suporte (funktionsverbgefüge) do português e do alemão*. In Cadernos do CIEG Centro Interuniver-sitário de Estudos Germanísticos. n. 1. Coimbra, Portugal: Universidade de Coimbra

Barreiro A. 2008. *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation*. Ph. D. thesis, Universidate do Porto

Biçici E and 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 (CICLing 2008), LNCS, Haifa, Israel, February 2008.

But M. 2003. *The Light Verb Jungle*. In Harvard Working Papers in Linguistics, ed. G. Aygen, C. Bowern, and C. Quinn. 1–49. Volume 9, Papers from the GSAS/Dudley House workshop on light verbs.

Dong M., Cheng Y., Liu Y., Xu J., Sun M., Izuha T., and Hao J. 2014. *Query lattice for translation retrieval*. In COLING.

Gupta R. and Orasan C. 2014. *Incorporating Paraphrasing in Translation Memory Matching and Retrieval*. In Proceedings of the 17th Annual Conference of European Association for Machine Translation.

He Y., Ma Y., van Genabith J., and Way A. 2010a. *Bridging SMT and TM with translation recommendation*. In ACL.

He Y., Ma Y., Way A., and Van Genabith J. 2010b. *Integrating n-best SMT outputs into a TM system*. In COLING.

Hirschberg Daniel S. 1997. *Serial computations of Levenshtein distances*. Pattern matching algorithms, Oxford University Press, Oxford.

Hodász G., & Pohl G. 2005. *MetaMorpho TM: a linguistically enriched translation memory*. In In international workshop, modern approaches in translation technologies.

Kay M. 1980. *The proper place of men and machines in language translation*. Palo Alto, CA: Xerox Palo Alto Research Center, October 1980; 21pp.

Koehn P. and Senellart J. 2010. *Convergence of translation memory and statistical machine translation*. In AMTA.

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

Pekar V., & Mitkov R. 2007. *New Generation Translation Memory: Content-Sensivite Matching*. In Proceedings of the 40th anniversary congress of the swiss association of translators, terminologists and interpreters.

Petrov S., Leon B., Romain T, and Dan K. 2006. *Learning accurate, compact, and interpretable tree annotation*. In Proceedings of the COLING/ ACL, pages 433–440.

Planas E., & Furuse O. 1999. *Formalizing Translation Memories*. In Proceedings of the 7th machine translation summit (pp. 331–339).

Scott B, Barreiro A 2009. *Openlogos MT and the SAL representation language*. In: Proceedings of the first international workshop on free/open-source rule-based machine translation, Alacant, pp 19–26

Silberztein M. 2003. *NooJ manual*. Available at http://www.nooj4nlp.net

Smith J. and Clark S. 2009. *Ebmt for SMT*: A new EBMT-SMT hybrid. In EBMT.

Wang K., Zong C., and Su K.-Y. 2014. *Dynamically integrating cross-domain translation memory into phrase-based machine translation during decoding*. In COLING.

Zhechev V. and van Genabith J. 2010. *Seeding statistical machine translation with translation memory output through tree-based structural alignment*. In SSST.