Translation Memory Systems Have a Long Way to Go

Andrea Silvestre Baquero[1], **Ruslan Mitkov**[2]
[1]Polytechnic University of Valencia, [2]University of Wolverhampton
andreasilvestre4@gmail.com, r.mitkov@wlv.ac.uk

Abstract

The TM memory systems changed the work of translators and now the translators not benefiting from these tools are a tiny minority. These tools operate on fuzzy (surface) matching mostly and cannot benefit from already translated texts which are synonymous to (or paraphrased versions of) the text to be translated. The match score is mostly based on character-string similarity, calculated through Levenshtein distance. The TM tools have difficulties with detecting similarities even in sentences which represent a minor revision of sentences already available in the translation memory. This shortcoming of the current TM systems was the subject of the present study and was empirically proven in the experiments we conducted. To this end, we compiled a small translation memory (English-Spanish) and applied several lexical and syntactic transformation rules to the source sentences with both English and Spanish being the source language.

The results of this study show that current TM systems have a long way to go and highlight the need for TM systems equipped with NLP capabilities which will offer the translator the advantage of he/she not having to translate a sentence again if an almost identical sentence has already been already translated.

1. Introduction

While automatic translation has taken off to work reasonably in some scenarios and to do well for gisting purposes, even today, against the background of the latest promising results delivered by statistical Machine Translation (MT) systems such as Google Translate and latest developments in Neural Machine Translation and in general Deep Learning for MT, automatic translation gets it often wrong and is not good enough for professional translation. Consequently, there has been a pressing need for a new generation of tools for professional translators to assist them reliably and speed up the translation process. Historically, it was Krollman who first put forward the reuse of existing human translations in 1971. A few years later, in 1979 Arthern went further and proposed the retrieval and reuse not only of identical text fragments (exact matches) but also of similar source sentences and their translations (fuzzy matches). It took another decade before the ideas sketched by Krollman and Arthern were commercialised as a result of the development of various computer-aided translation (CAT) tools such as Translation Memory (TM) systems in the early 1990s. These translation tools revolutionised the work of translators and the last two decades saw dramatic changes in the translation workflow.

The TM memory systems indeed changed the work of translators and now the translators not benefiting from these tools are a tiny minority. However, while these tools have proven to be very efficient for repetitive and voluminous texts, they operate on fuzzy (surface) matching mostly and cannot benefit from already translated texts which are synonymous to (or paraphrased versions of) the text to be translated. The match score is mostly based on character-string similarity, calculated through Levenshtein distance ("measure of the minimum number of insertions, deletions and substitutions needed to change one sequence of letters into another."; Somers 2003)

The limitations of the traditional TM systems in terms of matching have been highlighted by a number of authors. By way of example, Somers (2003) gives the example below to illustrate one drawback:

a. Select 'Symbol' in the Insert menu.
b. Select 'Symbol' in the Insert menu to enter character from the symbol set.
c. Select 'Paste' in the Edit menu.

The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT), pages 44–51,
Varna, Bulgaria, Sept 7, 2017
https://doi.org/10.26615/978-954-452-042-7_006

Given the input sentence (a), (c) would be the highest match percentage, as it differs in only two words, while (b) has eight additional words. Intuitively (b) is a better match since it includes the text of (a) in its entirety.

Before Mitkov (2005) elaborated on the specific matching limitations of TM systems, Macklovitch and Russel (2000) pointed out 'Current Translation Memory technology is limited by the rudimentary techniques employed for approximate matching'. They went on to illustrate that unless an TM system can do morphological analysis, it will have difficult recognising that (f) is more similar to (d) than (e):

d. The wild child is destroying his new toy
e. The wild chief is destroying his new tool
f. The wild children are destroying their new toy

Gow (2003) notes that SDL Trados gives the segments 'Prendre des mesures de dotation et de classification.' and 'Connaissance des techniques de rédaction et de révision.' a match rating of as high as 56% even though the above sentences have nothing to do with each other the reason being that because half of the words are the same and they are in the same position, even though the common words are only function words.

Recent work on new generation TM systems (Gupta 2015; Gupta et al. 2016a; Gupta et al. 2016b; Timonera and Mitkov 2015; Gupta and Orasan 2014) show that when NLP techniques such as paraphrasing or clause splitting are applied, TM systems performance is enhanced.

While that it is clear that TM systems are incapable of any linguistic or semantic interpretation, we maintain that they have difficulties with detecting similarities even in sentences which represent a minor revision of sentences already available in the translation memory. In order to substantiate this claim empirically, we conducted the following experiments.

We conducted experiments using a small translation memory (English-Spanish) in which we apply several lexical and syntactic transformation rules to the source sentences – both for English and Spanish serving as source language. The transformation rules are selected in such a way that they are simple and the transformed sentences do not change in meaning. The hypothesis of this study is that in many cases the TM systems cannot detect the fact that the transformed sentences are practically the same as sentences already translated as the match computed is below the threshold. This in turn, would mean insufficient efficiency as the new, practically the same sentences, will have to be translated again. For the purpose of this study we experimented with the well-known TM systems Trados, Wordfast, Omega T and MemoQ.

2. Data Description

For the purpose of this experiment a translation memory or alternatively parallel corpora were needed. To this end, we compiled a parallel English-Spanish corpus consisting of online documents of the European Union and the United Nations. The English documents selected were *Charter of the Fundamental Rights*, *Declaration on the Right to Development*, *Declaration on the Rights of Persons Belonging to National or Ethnic, and Religious and Linguistic Minorities*, *United Nations Declaration on the Rights of Indigenous People*, and *Universal Declaration of Human Rights*. We also selected the Spanish equivalents of these documents: *Carta de los derechos fundamentals de la Unión Europea*, *Declaración de las Naciones Unidades sobre el derecho al desarrollo*, *Declaración sobres los derechos de las personas pertenecientes a minorías nacionales o étnicas, religiosas y lingüísticas, Declaración de las Naciones Unidades sobre los derechos de los pueblos indígena* and *Declaración universal de los derechos humanos*.[1] The size of the English corpus was 14,153 words while the size of the Spanish corpus is 15,461 words (for more details see Table 1).

[1] The documents in English and Spanish are identical in contents even though the titles are slightly different.

English documents		Spanish documents	
Name	Size	Name	Size
1. Charter of the Fundamental Rights of the European Union	4,143	1. Carta de los derechos fundamentales de la Unión Europea	4,357
2. United Nations on the Right to Development	1,926	2. Declaración de las Naciones Unidas sobre el derecho al desarrollo	2,166
3. United Nation Declaration on the Rights of Indigenous Peoples	4,001	3. Declaración de las Naciones Unidas sobre los pueblos indígenas	4,427
4. Declaration on the Rights of Persons Belonging to National or Ethnic, Religious and Linguistic Minorities	2,309	4. Declaración sobre los derechos de las personas pertenecientes a minorías nacionales o étnicas, religiosas y lingüísticas	2,552
5. Universal Declaration of Human Rights	1,778	5. Declaración Universal de Derechos Humanos	1,959
Total documents : 5	Total words: 14,153	Total documents : 5	Total words: 15,461

Table 1: The experimental translation memory

We are aware the compiled corpus is very small but we regard this study and results as preliminary. In fact for the purpose of our experiments we selected a small sample of 150 aligned sentences in English and Spanish; this 'experimental TM' served as a basis for the experiments outlined.

3. Transformations

For the purpose of this study, we developed 10 transformation rules. Rule 1 was to transform an original sentence in active voice into a passive voice sentence, if possible. Rule 2 was a mirror image of rule 1 – transform a sentence in passive voice into active voice sentence. Rule 3 had to do with changing the order inside the sentence – changing the order of words, phrases or clauses within a sentence. Rule 4 sought to replace a word with a synonym whereas Rule 5 replaced 2 words of a sentence with their synonyms. The replacement with synonyms was applied to nouns, verbs and adverbs and in cases where the existence of synonym with identical meaning was 'obvious'. Rule 6 built on rule 3 by changing the order within a sentence but in addition replaced a noun with a pronoun for which it served as antecedent. Rule 7 was a combination of rule 1 and 4 – change of active voice into passive and replacement of a word with its synonym. Rule 8 changed passive voice into passive and like rule 6 - noun with a coreferential pronoun. Rule 9 was combination of rule 3 and rule 5, and finally was a subsequent application of rule 2, rule 3 and pronominalisation of a noun.

Table 2 below lists the rules with examples in English, whereas Table 3 lists the rules with examples in Spanish.

Rule	Transformation	Original sentence (English)	Transformed sentence (English)
1	Change active to passive voice	States must take measures to protect and promote the rights of minorities and their identity.	Measures must be taken to protect and promote the rights of minorities and their identity.
2	Change passive to active voice	The history, traditions and cultures of minorities must be reflected in education.	The education must reflect the history, traditions and cultures of minorities.
3	Change word order, phrase order or clause order	Reaffirming those indigenous peoples, in the exercise of their rights, should be free from discrimination of any kind.	Reaffirming those indigenous peoples should be free from discrimination of any kind, in the exercise of their rights.
4	Replace one word with its synonym	Everyone has the right to	Every person has the right to

		nationality.	nationality.
5	Replace two words with its synonym	Respect for the rights of the defence of anyone who has been changed shall be guaranteed.	Consideration for the rights of the protection of anyone who has been changed shall be guaranteed.
6	Replace one word into a pronoun AND change word order, phrase order or clause order	Indigenous peoples have the right to access, without any discrimination, to all social and health services.	They also have the right to access to all social and health services without any discrimination.
7	Change active to passive voice AND replace one word with its synonym	All children, whether born in or out of wedlock, shall enjoy the same social protection.	The identical social protection shall be enjoyed by all children, whether born in or out wedlock.
8	Change passive to active voice AND replace one word into a pronoun	No one shall be arbitrarily deprived of his property.	It shall not arbitrary deprive.
9	Change word order, phrase order or clause order AND replace two words with their synonyms	This booklet captures the essence of the Declaration, which is printed in full in this publication.	This brochure is printed in full in this publication which captures the nature of the Declaration.
10	Change active to passive voice AND change word order, phrase order or clause order AND replace one word into a pronoun	Enjoyment of these rights entails responsibilities and duties with regard to other persons, to human community and to future generations.	By enjoyment of these rights is involved the responsibilities and duties with regard to them, to the human community and to forthcoming generations.

Table 2: Transformation rules and examples in English

Rule	Transformation	Original sentence (Spanish)	Transformed sentence (Spanish)
1	Change active to passive voice	Destacando que corresponde a las Naciones Unidas desempeñar un papel importante y continuo de promoción y protección de los derechos de los pueblos indígenas.	Destacando que es correspondido por las Naciones Unidades desempeñar un papel importante y continuo de promoción y protección de los derechos de los pueblos indígenas.
2	Change passive to active voice	El contacto pacífico entre minorías no debe ser restringido.	Los Estados no deben restringir el contacto pacífico entre minorías.
3	Change word order, phrase order or clause order	Los Estados, sin perjuicio de la obligación de expresión, deberán alentar a los medios de información privados a reflejar debidamente la diversidad cultural indígena.	Los Estados deberán alentar a los medios de información privados a reflejar debidamente la diversidad indígena, sin perjuicio de la obligación de asegurar plenamente la libertad de expresión.
4	Replace one word with its synonym	Se garantiza la protección de la familia en los planos jurídico, económico y social.	Se asegura la protección de la familia en los planos jurídico, económico y social.
5	Replace two words with its synonym	Las sociedades de todo el mundo disfrutan de la diversidad étnica, lingüística y religiosa.	Las sociedades mundiales disfrutan de la variedad étnica, lingüística y religiosa.
6	Replace one word into a pronoun AND change word order, phrase order or clause order	Todos los niños, nacidos de matrimonio o fuera de matrimonio, tienen derecho a igual protección social.	Ellos tienen derechos a igual protección social, nacidos de matrimonio o fuera de matrimonio.
7	Change active to passive voice AND replace one word with its synonym	Los Estados tomarán las medidas que sean necesarias para lograr progresivamente que este derecho se haga plenamente efectivo.	Las decisiones que sean necesarias para lograr progresivamente que este derecho se haga plenamente efectivo serán tomadas por los Estados.

8	Change passive to active voice AND replace one word into a pronoun	La igualdad entre hombres y mujeres será garantizada en todos los ámbitos, inclusive en materia de empleo, trabajo y retribución.	Ellos garantizarán la igualdad entre hombres y mujeres en todos los ámbitos, inclusive en materia de empleo, trabajo y retribución.
9	Change word order, phrase order or clause order AND replace two words with their synonyms	Del ejercicio de ese derecho no puede resultar discriminación de ningún tipo.	No puede surgir discriminación de ningún tipo de la actuación del ejercicio de ese derecho.
10	Change active to passive voice AND change word order, phrase order or clause order AND replace one word into a pronoun	Al definirse y ejecutarse todas las políticas y acciones de la Unión se garantizará un alto nivel de protección de la salud humana.	Un alto nivel de protección de la salud humana será garantizada al ejecutarse y definirse todas las políticas y acciones de ella.

Table 3: Transformation rules and examples in Spanish

4. Experiments, Results and Discussion

We use the above parallel corpus as a translation memory and experiment with both English and Spanish as source languages. If we had to translate again a sentence from the source Language, the match would be obviously 100%. For the purpose of the experiment, each sentence of the source text undergoes a transformation after applying the rules listed below, which convert the original sentences into syntactically different but semantically same sentences. As the new sentences have the same meaning, it would be desirable that the TM systems produce a high match between the transformed sentences and the original ones. By 'high match' we mean a match above the threshold of a specific TM tool so that the user can benefit from the translation of the original sentence being displayed.

Rule	# Sentences	Trados		Wordfast		OmegaT		MemoQ	
		< 75%	Failure %	<75%	Failure%	<75%	Failure%	<75%	Failure%
1	28	10	35.71	2	7.14	14	50	11	39.28
2	27	7	25.92	6	22.2	14	51.85	13	48.15
3	84	8	9.52	0	0	18	21.43	24	28.57
4	150	2	1.33	21	14	6	4	13	8.6
5	150	10	6.67	77	51	7	4.6	62	41.3
6	29	11	37.93	5	17.24	14	48.27	16	55.17
7	26	9	34.61	14	53.85	11	42.31	22	84.61
8	9	2	22.22	3	33.3	5	55.55	6	66.67
9	84	22	26.19	42	50	42	50	64	76.2
10	20	4	20	5	25	9	45	14	70

Table 4: Matching results English

The results obtained with English as a source language show that the lexical and syntactic transformations cause significant problems to the TM systems and their inability to return matching above the default threshold means that the translator may miss out on the re-use of translations he/she has already made. In the case of Trados there are as many as 36% of the sentences not being returned after being transformed with rule 1; this figure goes up to 38% when applying rule 6 and 35% with rule 7. Wordfast fails to propose similarity for 54% of the sentences once rule 7 is applied. As for Omega T, the application of rule 8 results in

56% of the transformed sentences not being detected as similar to the original sentence. MemoQ's failures are even more dramatic in that the applications of rules 10, 9 and 7 leads to inability to detected similarity in 70%, 76% and 85% of the cases respectively! Overall, MemoQ reports the most negative results which is surprising given the very positive feedback of users in general for this tool.
It appears that TM systems in general have more problems with syntactic transformations – after applying rules 1, 2, 3 or the combined rules 6-10. The TM systems also report fewer problems for lexical transformation only – e.g. after applying rules 4 and 5 only.

As the transformations are language specific, we conducted similar experiments with Spanish being the source languages. We transformed the original Spanish sentences using the above rules and the TM systems computed the match between the transformed sentences and the original ones.

Rule	# Sentences	Trados		Wordfast		OmegaT		MemoQ	
		<75%	Failure %	<75%	Failure%	<75%	Failure%	#<75%	Failure%
1	50	10	20	7	14	22	44	12	24
2	18	11	61.1	10	55.55	11	61.11	12	66.67
3	91	20	21.98	1	1.10	19	20.88	23	25.27
4	150	5	3.33	24	16	5	5	14	9.33
5	150	13	8.67	95	63.3	33	22	55	36.67
6	44	18	40.9	5	11.36	16	36.36	20	45.45
7	50	16	32	29	58	19	38	24	48
8	17	10	58.82	8	47.05	10	58.82	9	52.94
9	91	36	39.56	50	54.94	41	45.05	55	60.44
10	25	6	24	9	36	9	36	18	72

Table 5: Matching results Spanish

The results obtained with Spanish as a source language show that the lexical and syntactic transformations cause more significant problems to the TM systems than with English as a source language. In the case of Trados there are as many as 40% of the sentences not being returned after being transformed with rule 6 and 9; this figure goes up around 60% when applying rule 8 and 61% with rule 2. Wordfast fails to propose similarity for around 55% of the sentences once rule 2 and 9 are applied; and up to 63% when rule 5 is also applied. As for Omega T, the application of rule 8 results in 59% of the transformed sentences not being detected as similar to the original sentence and up to 61% when applying rule 2. MemoQ's failures after applying rules 8, 9 and 2 leads to inability of detecting similarity in 52%, 60% and 67% of the different cases. On the whole, every TM system exhibits retrieval failures mostly related to combining different rules.

It is worth noting the high errors rates for all TM systems even with rule 2 which confirms their inability to deal with syntactic transformations.

Finally, it is worth observing that for Spanish as source language the matching failures are higher. As the above TM systems have no NLP functionalities and usually use Levenstein distance as a matching algorithm, we conjecture that this has to do with the slightly more complex syntax in Spanish.

5. Conclusion

Current TM systems have a long way to go. The above results highlight the need for TM

technology to embrace NLP techniques such as parsing or paraphrasing. A TM system equipped with NLP capabilities will offer the translator the advantage of he/she not having to translate a sentence again if an almost identical sentence has already been already translated.

References

Arthern, Peter. 1979. “Machine translation and computerised terminology systems: a translator's viewpoint.” Edited by Barbara Snell. Translating and the computer Amsterdam, North-Holland. 77-108.

Gow, Francie. 2003. “Metrics for Evaluating Translation Memory Software.” MA thesis, University of Ottawa, Canada.

Grönroos, Mickel, and Ari Becks. 2005. “Bringing Intelligence to Translation Memory Technology.” Proceedings of the International Conference Translating and the Computer 27. London: ASLIB.

Gupta, R. 2015. *Use of Language technology to improve matching and retrieval in Translation Memory*. PhD thesis. University of Wolverampton.

Gupta, R., Orasan, C., Zampieri, M., Vela. M., Mihaela Vela, van Genabith, J. and R. Mitkov. 2016a. "Improving Translation Memory matching and retrieval using paraphrases", *Machine Translation*, 30(1), 19-40.

Gupta, R., Orasan, C., Liu, Q. and R. Mitkov. 2016b. A Dynamic Programming Approach to Improving Translation Memory Matching and Retrieval using Paraphrases. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue (TSD)*, Brno, Czech Republic.

Gupta, R. and Orasan, C. 2014. Incorporating Paraphrasing in Translation Memory Matching and Retrieval. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014),* 3-10. Dubrovnik, Croatia.

Gupta, R., Bechara, H. and C. Orasan. 2014. "Intelligent Translation Memory Matching and Retrieval Metric Exploiting Linguistic Technology". *Proceedings of the Translating and Computer 36*, 86-89.

Hodász, Gábor, and Gábor Pohl. 2005. “MetaMorpho TM: a linguistically enriched translation memory.” Edited by Walter Hahn, John Hutchins and Cristina Vertan. International Workshop, Modern Approaches in Translation Technologies..

Hutchins, John. 1998. “The origins of the translator’s workstation.” Machine Translation 13, n° 4: 287-307.

Kay, Martin. 1980. “The Proper Place of Men and Machines in Language Translation.” Machine Translation 12, n° 1-2: 3-23.

Lagoudaki, Pelagia Maria. 2008. “Expanding the Possibilities of Translation Memory Systems: From the Translator’s Wishlist to the Developer’s Design.” PhD diss., Imperial College of London

Lagoudaki, Pelagia Maria. 2006. “Translation Memories Survey 2006: Users’ perceptions around TM use.” Translating and the Computer 28 (ASLIB).

Macklovitch, Elliott, and Graham Russell. 2000. “What's Been Forgotten in Translation Memory.” AMTA '00 Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future. London: Springer-Verlag.137-146.

Marsye, Aurora. 2011. “Towards a New Generation Translation Memory: A Paraphrase Recognition Study for Translation Memory System Development.” Master’s Thesis. University of Wolverhampton and Université de Franche-Comté..

Mitkov, Ruslan, and Gloria Corpas. 2008. “Improving Third Generation Translation Memory systems through identification of rhetorical predicates.” Proceedings of the LangTech 2008 conference. Rome,

Mitkov, Ruslan. 2005. ‘New Generation Translation Memory systems’. Panel discussion at the 27th international Aslib conference ‘Translating and the Computer’. London

Pekar, Viktor, and Ruslan Mitkov. 2007. “New Generation Translation Memory: Content-Sensitive Matching.” Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters. Bern: ASTTI

Nicolas, Lionel, Egon Stemle, Klara Kranebitter and Verena Lyding. 2013. “ High-Accuracy Phrase Translation Acquisition through Battle-Royale Selection “. Proceedings of RANLP’2013.

Planas, Emmanuel, and Osamu Furuse. 1999. “Formalizing Translation Memories.” Proc MT Summit VII. 331-339.

Planas, Emmanuel. 2005. “SIMILIS: Second-generation translation memory software.” proceedings of the 27th International Conference Translating and the Computer. London: Reinke, Uwe. State of the Art in Translation Memory Technology. 2013. *Translation: Computation, Corpora, Cognition*, [S.l.], v. 3, n. 1, jun. 2013. ISSN 2193-6986.

Somers, Harold. “Translation Memory Systems.” 2003. In Computers and Translation: A Translator's Guide, Edited by Harold Somers,

31-47. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Steinberger, Ralf, et al. 2006. "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages." Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy.

Timonera, K. and R. Mitkov. 2015. Improving Translation Memory Matching through Clause Splitting. Proceedings of the RANLP'2015 workshop 'Natural Language Processing for Translation Memories'. Hissar, Bulgaria.

Zhechev, Ventsislav and Josef van Genabith. 2010. Maximising TM Performance through Sub-Tree Alignment and SMT. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*