# Improving Third Generation Translation Memory Systems Through Identification of Rhetorical Predicates

*Ruslan Mitkov* [1], *Gloria Corpas* [2]

[1] University of Wolverhampton
[2] University of Malaga
r.mitkov@wlv.ac.uk, gcorpas@uma.es

## Abstract

While number of Translation Memory (TM) programs and tools have been developed which are now regarded as indispensable for the work of professional translators, it has been noted that a serious weakness of the current TM technology is the fact that its matching capability is far from perfect. An obvious shortcoming of current TM systems is the fact that they have no access to the meaning of the translated text and operate on its surface form. As a result, they fail to match sentences that have the same meaning, but different syntactic structure. To overcome this shortcoming Pekar and Mitkov (2007) developed the so-called 3rd Generation Translation Memory (3GTM) methodology which analyses the segments not only in terms of syntax but also in terms of semantics. Whereas this technology is a promising way forward, the limitations of current semantic processing may cast a doubt on its use in a practical environment. To enhance the overall low performance of semantic processing tasks, we propose the employment of rhetorical predicates to improve the accuracy of the matching algorithm. The paper will introduce the novel 3GTM developed by us and will show how rhetorical predicates can be used to enhance its performance.

**Index Terms:** Translation Memory, 3rd Generation Translation Memory, Translation Memory, matching algorithm, semantic processing, rhetorical predicates, Natural Language Processing.

## 1. Shortcomings of traditional Translation Memory systems and the development of 3rd Generation Translation Methodology

Translation Memory (TM) systems have emerged as highly successful translation aids for more than a decade. TM has proven to be efficient and time-saving technology for voluminous translations especially of technical texts featuring a degree of repetition with previously translated texts. While a number of TM programs and tools have been developed which are now regarded as indispensable for the work of professional translators, it has been noted that a serious weakness of the current TM technology is the fact that its matching capability is far from perfect.

As noted by Mitkov (2005), traditional TM systems would fail to return matches for parts of sentences which would match on their own. By way of example, if *Select 'Shut down'* has been previously translated, that a translation of *Select 'Shut down' from the menu* would not return any partial matches. Similarly, a complex sentence consisting of two previously matched simple sentences, would not offer any matches. As an illustration, if *Select 'Shut down' from the menu* and *Click on 'Shut down'* match, *Select 'Shut down' from the menu and click on 'Shut down'* would not be returned as a match.

A more serious shortcoming of current TM systems is the fact that they have no access to the meaning of the translated text and operate on its surface form. As a result, they fail to match sentences that have the same meaning, but different syntactic structure. The fact that the same semantics can be expressed in a variety of linguistic forms has a number of important consequences for the practical use of TM systems. By way of example, none of the TM systems would be capable of matching *Microsoft developed Windows XP* with *Windows XP was developed by Microsoft* or matching *The company bought shares* with *The company completed acquisition of shares*.

To overcome this shortcoming Pekar and Mitkov (2007) developed the so-called *3rd Generation Translation Memory* (3GTM) methodology[1] which analyses the segments not only in terms of syntax but also in terms of semantics. The authors adopt the so-

[1] SIMILIS, developed by Lingua et Machina (Planas, 2005), was referred to as "Second generation TM software". It does morpho-syntactic analysis (chunks) and holds translation units corresponding to chunks: e.g., as with parallel concordancers, matching at sub-sentence level is possible. By way of example in the translation of 'Les mémoires de seconde génération changent le monde de la traduction' into 'Second generation memories change the translation world', the match of the noun phrases *mémoires de seconde generation* and *second generation memories* is possible. Similar work has been reported in Grönroos and Becks (2005) and in Hodasz and Pohl (2005).

called syntax-driven semantic analysis by performing linguistic processing over trees graphs (Graehl and Knight 2004; Szpektor et al. 2004) followed by lexico-syntactic normalisation. Then similarity between syntactic-semantic tree graphs is computed and matches at sub-sentence level are established, using a similarity filter and node distance filter (Pekar and Mitkov 2007).

Phase 1 of this project covers the implementation and evaluation of the new matching technology through the development of a matching algorithm including the pre-processing modules, the integration with WordNet and the development of a lexical paraphrase module. Phase 2 includes the integration within Wordfast and a translator's evaluation.

Lexical resources such as WordNet are needed to identify automatically synonyms and therefore, to make the match of synonymous expressions possible. By way of example, if 'Unplug the hoover' has already been translated, then of 'Unplug the vacuum cleaner' would show as match too as *vacuum cleaner* would be labelled as a synonym of *hoover*. In addition, generalisations would be possible as WordNet would return *machine* as a superordinate of *vacuum cleaner* so a high match for 'Unplug the machine' would be returned. Similar generalisations would be achieved for patterns with identical or seminal semantic classes such as 'John Smith few to Brussels on February 3rd' or 'Dr Johnson flew to Rome on 7 January 2006' or even more generally '<person-m> flew to <city> on <date>'. Finally a lexical paraphrase resource is made use of to establish equivalences between expressions such as 'Microsoft developed Windows XP' vs. 'Windows XP was developed by Microsoft' or 'The company bought shares' vs. 'The company completed acquisition of shares'. For more complicated cases such as matching 'As a result of John Smith's resignation, the values of the company shares plummeted' with 'John Smith resigned and this resulted in a sharp decrease of the values of the company shares', textual entailment techniques may be needed.

## 2. Limitations of the 3rd Generation Translation Memory Methodology

The 3GTM methodology proposed by Pekar and Mitkov (2007) is a promising way forward to ensure that translators have wider range of matches. However, this methodology is still a long way from operating in a practical environment and in particular, from being commercialised. A major issue is whether the new technology would deliver in a robust and scalable way. Even though contingency plans include the implementation of a subset of the techniques, Natural Language Processing (NLP) in general is far from perfect. With particular reference to the semantic tasks involved, the accuracy could be as low as 60% for semantic role labelling or even lower for anaphora resolution (see the last example from the previous paragraph). Therefore, different techniques should be sought to enhance the success rate of semantic processing.

## 3. A novel rhetorical predicates-driven methodology for 3rd Generation Translation Memory systems

Our presentation will discuss a new project whose objective is to enhance the matching performance when comparing semantically equivalent sentences through the identification of rhetorical predicates[1]. Research in discourse, text generation and text summarisation has already shown that different types of texts feature schemata of rhetorical predicates which account for the stereotypical discourse structure of the specific type of text. By way of example, a research paper normally features among others, rhetorical predicates such as *topic*, *background*, *methodology*, *solution* and *conclusion*. While such a list of rhetorical predicates is genre- or domain-specific, we argue that the identification of rhetorical predicates will assist the establishment of equivalence of sentences. The methodology consists of boosting the confidence/probability that two sentences are semantically equivalent if these sentences are labelled with the same rhetorical predicates or vice versa. The identification of rhetorical predicates will be carried out on the basis of regular expressions containing keywords representative of a specific predicate.

The presentation as well as the final version of the paper will introduce the novel 3GTM methodology developed by us, describe the above experiments and will report on the evaluation of the matching performance of a 3rd Generation Translation Memory system – with and without a module for identification of rhetorical predicates. The experiments will be restricted to specific genres due to the nature of rhetorical predicates. As a convenient background, a comparative assessment of current MT systems (e.g. Déjà vu, Trados, Wordfast, etc.) performance will be also provided. We intend to give the presentation in an easy-to-follow and accessible manner which will be based more on translation/matching examples rather than technicalities, while elaborating on technical issues in the paper itself.

References

[1] Graehl J. and Knight K. 2004. "Training Tree Transducers". *Proceedings of NAACL/HLT-2004*. Boston, MA.

[2] Hodasz G. and Pohl G. 2005. "MetaMorpho TM: A Linguistically Enriched Translation Memory". *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-05).* Borovets, Bulgaria.

[1] Termed also *moves*.

[3] Mitkov, R. 2005. “Panel Discussion: The Future of TM Technology”. *27th International Conference on Translating and the Computer (TC27)*. London, UK.

[4] Pekar V. and Mitkov R. 2007. “New Generation Translation Memory: Content-Sensitive Matching”. *Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters*, 29-30 September 2006, Bern.

[5] Planas, E. 2005. “SIMILIS - Second generation TM software”. *Proceedings of the 27th International Conference on Translating and the Computer (TC27)*. London, UK.