A Virtuous Circle: Laundering Translation Memory Data using Statistical Machine Translation. Session 3 - Machine and Human Translation: Finding the Fit?

Joss Moorkens, Stephen Doherty, Dorothy Kenny, Sharon O'Brien

TRALOGY

A Virtuous Circle: Laundering Translation Memory Data using Statistical Machine Translation

Joss Moorkens
Centre for Next Generation Localisation, Centre for Translation & Textual Studies, Dublin City University
joss.moorkens@dcu.ie

Stephen Doherty
Centre for Next Generation Localisation, Centre for Translation & Textual Studies, Dublin City University
stephen.doherty@dcu.ie

Dorothy Kenny
Centre for Next Generation Localisation, Centre for Translation & Textual Studies, Dublin City University
dorothy.kenny@dcu.ie

Sharon O'Brien
Centre for Next Generation Localisation, Centre for Translation & Textual Studies, Dublin City University
sharon.obrien@dcu.ie

TRALOGY II - Session 3
Date d'intervention : 17/01/2013

This study compares consistency in target texts produced using Translation Memory (TM) with that of target texts produced using Statistical Machine Translation (SMT), where the SMT engine is trained on the same texts as are reused in the TM workflow. These comparisons focus specifically on noun and verb inconsistencies, as such inconsistencies appear to be highly prevalent in TM data (Moorkens 2012). We then go on to substitute inconsistent TM target text nouns and verbs by consistent nouns and verbs from the SMT output, and to test (1) whether this results in improvements in overall TM consistency and (2) whether an SMT engine trained on the 'laundered' TM data performs better than the baseline engine. Improvements were observed in both TM consistency and SMT performance, a finding that indicates the potential of this approach to TM/MT integration.

lien video : http://webcast.in2p3.fr/videos-a_virtuous_circle

Introduction

While the functionality of TM tools has improved in the 20 years since their introduction, several studies have shown translation consistency to be a continuing issue (e.g. Rieche 2004, Moorkens 2012). Nevertheless, a core assumption behind the use of TM tools is that they can minimise inconsistency. In the past decade or so, there has also been an increasing move towards the use of translation memories in training statistical machine translation (SMT) engines, which raises the question: What impact does the use of inconsistent TM data have on SMT output? This study seeks to address this question by training an SMT engine on TM data whose inconsistencies have been analysed and documented in detail (Moorkens 2012). We also test whether, by making the TM more consistent, the SMT output improves. There has been prior work on TM and MT integration. Some such work investigates methods for deciding on TM fuzzy match thresholds below which it is better to present a translator with a machine translated segment rather than a fuzzy match from TM memory (O'Brien 2006, Guerberof 2012). Other work (e.g. He 2012) looks at how TM and MT outputs can be ranked, so as to present the best translation proposal to a translator. These sources are essentially concerned with TM interfaces and what is proposed to human translators as they translate. At sub-segment level, He (ibid.) also investigates whether phrase pairs derived from fuzzy TM matches can be used to constrain translations output by SMT. Our research is more concerned with the data used to train SMT systems before run-time. It is thus related both to He's (ibid.) work and to other data-oriented research carried out by Ozdowska and Way (2009).

The specific aims of this study are thus twofold; firstly, to discover whether the consistency of a TM may be significantly improved by substituting inconsistent nouns and verbs with consistent nouns and verbs from SMT output trained on that same TM. The second aim is to discover what impact any observed improvement in TM consistency has on the quality of output from an SMT engine trained on the more consistent TM data. This paper contains the initial results of the study using an English-to-German TM.

1. Methodology

1.1 Research Phases

The research project was carried out in three phases (see Figure 1 and below):

- 1. Identification of inconsistencies in the untreated TM and comparison with baseline SMT engine output;
- 2. Replacement of inconsistent nouns and verbs by those from the SMT output, thus creating a 'cleaned' (more consistent) TM, which was then used to create a 'cleaned' SMT engine;
- 3a. Comparison of the untreated TM and the 'cleaned' TM using the Xbench QA tool;
- 3b. Quality assessment of the output from both SMT engines using Automatic Evaluation Metrics (AEMs).

Results for phases 1 and 3 are presented in the Results section (figure 1.).

1.2 Data

The baseline data used in the current study consist of 22,691 TUs from an English-to-German TM, made up of general technical support documentation, supplied in the TMX format. The documentation was produced by a leading multinational software company in 2008. The file containing these data was named ENDE_Baseline for the purpose of comparison with later

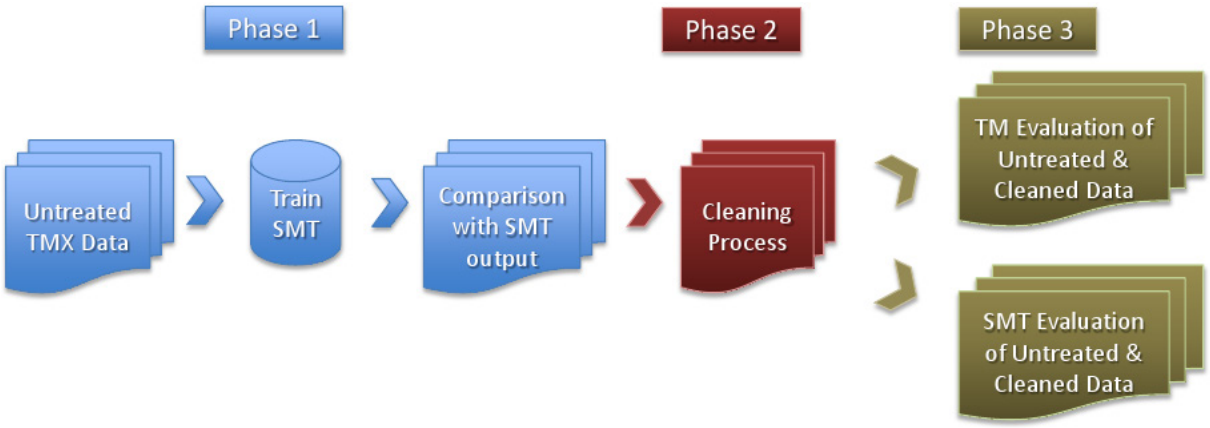

Figure 1: Three phases of the current study

versions of the TM. Once inconsistencies had been identified in the baseline TM data, they were replaced with consistent SMT output, using the UltraEdit editor. The 'cleaned' version of the TM was named ENDE_Cleaned and subsequently used as training data for the Smart MATE SMT engine.

1.3 Analysis of Inconsistencies

This work follows on from a quantitative study in Moorkens (2012) in which inconsistencies were identified at segment and sub-segment level. In Moorkens (ibid.) target text segments were described as 'inconsistent' if they differed when we could reasonably expect them to have been formally identical. We expect TT segments to be formally identical if the corresponding ST segments are formally identical or contain only minor non-semantic differences. In such cases, we talk of 'repeated' ST segments. In Moorkens (ibid.), sub-segment inconsistencies were largely categorised based on part-of-speech, such as noun, verb, or adverb. (Other categories in Moorkens (ibid.) included punctuation and word order inconsistencies.) Where there were more than three sub-segment inconsistencies within a single segment, the segment was deemed to have been completely rewritten. We reuse Moorkens's analytical categories in the current paper.

1.4 Consistency checking in TMs

As already indicated, in Phase 2, inconsistent nouns and verbs from inconsistent target segments were manually identified and replaced with consistent nouns and verbs from the SMT output, and the new TMX file saved as 'ENDE_Cleaned'. This file and the baseline TMX file were then tested using the automated QA check in Xbench for target segment consistency and the results compared. This was done to validate the cleaning phase and to confirm that the automated tool measured improved consistency within the TM.

1.5 Machine Translation System

The SMT system used in this study, SmartMATE (Way et al. 2011), is a cloud-based self-serve translation platform where users can upload content, e.g. TMX files, to create customized SMT engines. The system has proven to be successful and reliable in other related tasks (e.g. Doherty et al. 2012). The system adopts a predominantly statistical approach to the MT process (e.g. Koehn 2010). SMT works on the premise of learning from previously translated segments provided by human translators, typically by means of TMX files. Using these data, the engine constructs a monolingual language model for the source and target languages, and a bi-directional

translation model, in our case, from U.S. English into German. At runtime, it uses probabilities to select the most likely translation from a list of candidates (for a full description of SMT, see Hearne and Way 2011).

1.6 Training and Evaluating the SMT Engines

In order to test whether training an SMT engine with more consistent TM data would result in better SMT output, one test set of 20 translation units was removed from the baseline TM (in phase 1) and a further test set of 20 translation units was removed from the cleaned TM (in phase 2). The two test sets were based on the same source segments. (Recall that changes had been made only to the target-language side of the translation units in phase 2). Two SMT engines were thus trained; one using data from the original baseline TM, the other using data from the cleaned TM. In neither case were the test data included in the training data. The 20 source segments from the test set were then translated using the SMT engines trained on the baseline (phase 1) and cleaned TMX files (phase 2) and the results compared using AEMs. The stages of training and evaluation of the SMT engines used in the current study are summarised in Figure 3.

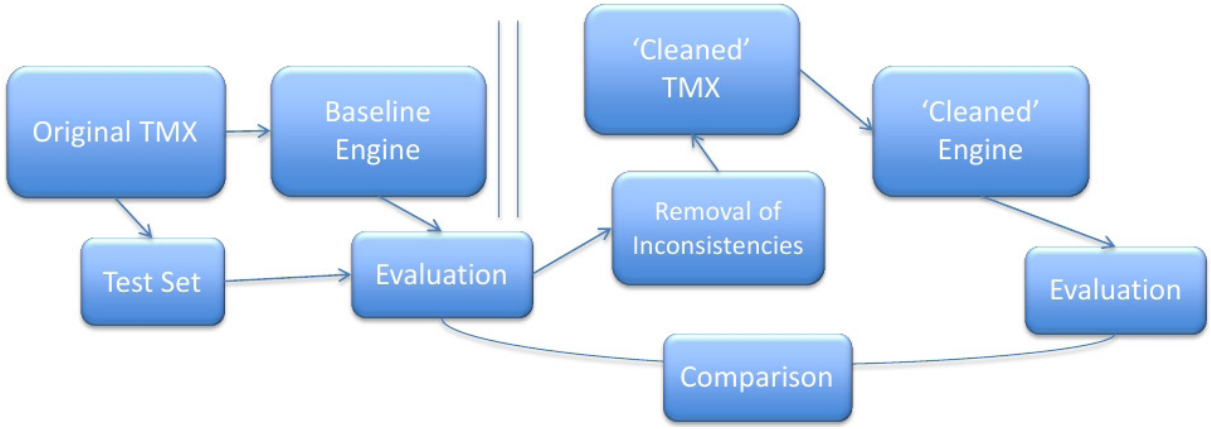

Figure 2: SMT training and evaluation

ApSIC Xbench was used to test for target segment inconsistency in the TM (Figure 1, phase 3). As part of its QA functionality, Xbench can run a check on a TM, listing categories such as: inconsistency in source, inconsistency in target, double spaces, and key term mismatch. While the tool is limited in that it only works at segment level, in this study it found as many identified consistencies as the manual method outlined in Moorkens (2012). It is widely used in the translation industry and has been shown to be effective in identifying translation inconsistencies (Debove, Furlan, and Depraetere, 2011 p185). Xbench also provides an alternative, automated method to the time-consuming manual identification of inconsistencies.

To evaluate SMT output, we used several popular automatic evaluation metrics: BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002), General Text Matcher (GTM) (Turian et al. 2003, Melamed et al. 2003), and Translation Edit Rate (TER) (Snover et al. 2006) to represent different aspects of the measure of inconsistency vis-à-vis textual similarity and difference.

BLEU counts the number of words that are found to be common to both a reference (gold standard human translation) and the MT engine output, and uses precision to compute the level of similarity. Scores range from 0 to 1, where 1 denotes that the two sentences are

identical, and 0 denotes no words in common. Similarly, GTM counts the number of common words between the MT output and the reference text, but also uses a recall measure to compute the level of similarity. Its score also ranges from 0 to 1, rising with increasing textual similarity. Lastly, TER counts the minim um number of edits necessary to transform the MT output into the reference text. Such edits can involve: insertions, deletions, substitutions, and changes in word order. Where no edits have been implemented, a score of 0 is given; and the greater the number of edits, the higher the score. The rationale for choosing these metrics is that they can be used with European languages and Japanese (the latter is also a focus of the current research project but is not reported here), and they are commonly used metrics in the research and development of TM and MT technologies in research and commercial settings.

2. Results

2.1 Phase 1

Analysis of target text inconsistencies in the baseline TM found 215 inconsistencies contained within 176 segments. 47% (101) of sub-segment target text inconsistencies in this TM were accounted for by noun inconsistency; and just below 20% (42) of target text inconsistencies by verb inconsistencies. In total, 66% of 215 sub-segment inconsistencies in this TM were noun and verb inconsistencies – see Figure 2. (By way of comparison, on average 60% of sub-segment inconsistencies were accounted for by noun and verb inconsistencies in the four comparable TMs studied in Moorkens 2012.) Thus reducing noun and verb inconsistencies should have an appreciable effect on the level of inconsistency within this TM.

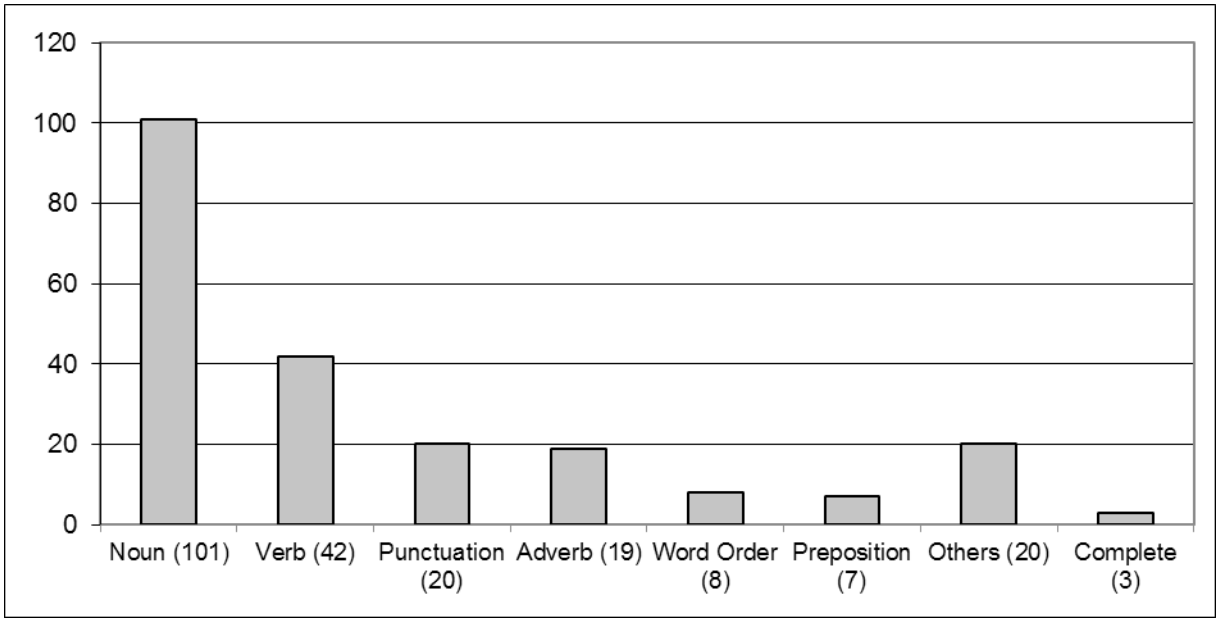

Figure 3: Target text inconsistencies found in TM used in current study

In Table 1, inconsistencies found manually in target segments in the baseline TM are compared with inconsistencies found in target segments produced by the baseline SMT engine. The final column records new errors introduced by the SMT engine.

Table 1: Target text inconsistencies in TM and SMT output

Category of TM inconsistency	Inconsistencies in target segments of baseline TM	Inconsistencies removed in SMT target segments	Errors added in SMT target segments
Noun	101	92	11 (e.g. 6 letter case, 3 nouns missing)
Verb	42	32	18 (e.g. 6 verb missing)
Adverb	19	13	7
Punctuation	20	14	1
Preposition	7	4	2
Article	6	3	0
Word Order	8	3	20
Tags	7	3	28
Typo	5	3	0
Completely rewritten	3	0	10
Total Sub-segment Inconsistencies	215	167 removed	87 added

As shown in Table 1, target segment inconsistency is reduced in SMT output, particularly for nouns and verbs, which are the most commonly-occurring categories of TM target segment inconsistency (Moorkens 2012). 92 of 101 noun inconsistencies (91%) were made consistent in the SMT output. Also, 32 of 42 verb inconsistencies (76%) were made consistent in the SMT output. However, 11 noun errors and 18 verb errors were introduced. These errors involve missing verbs (in 6 cases), German nouns that do not begin with a capital letter (in 6 cases), and missing nouns (in 3 cases). For example, in two cases the noun 'Gitterausrichtung' was translated as 'gitterausrichtung' (without capitalisation) by the SMT. Similarly, '3D-Schneidewerkzeug' became '3D-schneidewerkzeug'. Due to a typographical error in the TM, the word 'Häkchen' was written twice as 'Häckchen', and the incorrect version was used in the SMT output. Assuming we were to replace the inconsistent nouns and verbs with those from the SMT output, 124 inconsistencies would be removed and 29 errors potentially introduced. By replacing the inconsistent nouns and verbs with those from the SMT output, it would appear that total target text inconsistency could be reduced by 58% (124 inconsistencies removed) with 29 errors added.

2.2 Phase 3

In Phase 1, our manual analysis of the baseline TM found 176 inconsistent target segments, containing 215 sub-segment inconsistencies. When the baseline TM was tested again with the Xbench QA tool for the purposes of validation, 176 segment level TT inconsistencies were again found. A subsequent analysis in phase 3 of the 'cleaned' TM file found 80 target segment inconsistencies, showing that the replacement of inconsistent nouns and verbs had improved the consistency of the TM file as measured using a commercial QA tool. Looking more closely at the TMX file, however, some problems became apparent. Errors, such as those of letter case noted previously, were propagated in the SMT output. The SMT engine, for its part, was not always consistent in its translation of given nouns. For example, the TMX file contains two translations for 'dialog box': 'Dialogfeld' and 'Dialogfenster'. Both options are produced by the SMT engine as may be seen in Examples 1 and 2:

Example 1

Source Text	SMT Output
Click {1}Options…{2} in the {3}3D Projection{4} dialog box.	Klicken Sie auf Optionen…} 2 {1}{im Dialogfenster 3D Projektion {3}{4}.

Example 2

Source Text	SMT Output
Click {1}OK{2} to confirm your entry or {3} Cancel{4} to close the dialog box without making any changes.	Klicken Sie auf OK, um Ihre Eingabe zu bestätigen, oder {1}{2} {3}Abbrechen{4} das Dialogfeld ohne eventuelle Änderungen zu schließen.

In this study we have focussed on target segment inconsistency. The SMT output appears to give consistent results for nouns and verbs as long as the source text is consistent. While not part of this study, source text inconsistencies may be seen to lead to further inconsistency in the target text. For example, the TM contains three different TT translations of 'Placed Files Window'. These are: 'Platzierte Datei-Fenster', 'Fenster "Platzierte Datei"', and 'Fenster "Platzierte Dateien"'. The SMT translates the ST segment only as 'Platzierte Datei-Fenster', making the TT consistent. A problem arises when we compare the various translations of the ST segments 'Placed Files window' (with a lower case W) and 'Placed files window', as per Example 3:

Example 3: Three letter case variants of a ST segment and aligned target segments

Source Text	TM Target Text	SMT Target Text
Placed Files Window	Platzierte Datei-Fenster	Platzierte Datei-Fenster
Placed Files Window	Fenster «Platzierte Datei»	Platzierte Datei-Fenster
Placed Files Window	Fenster «Platzierte Datei»	Platzierte Datei-Fenster
Placed Files Window	Fenster «Platzierte Dateien»	Platzierte Datei-Fenster
Placed Files window	Fenster «Platzierte Datei»	Fenster «Platzierte Datei»
Placed Files window	Fenster «Platzierte Dateien»	Fenster «Platzierte Datei»
Placed Files window	Fenster «Platzierte Datei»	Fenster «Platzierte Dateien»
Placed Files window	Fenster «Platzierte Dateien»	Fenster «Platzierte Dateien»

The inconsistent ST letter case affects the SMT output, giving inconsistent results. The baseline TM contains 200 examples of inconsistent ST letter case (41% of total ST inconsistencies). Were these ST inconsistencies to be removed before translation, the SMT results (and possibly the human translation results as 60 of these ST inconsistencies are aligned with inconsistent TT) would be rendered more consistent. This will be tested in a later iteration of this study.

2.3 AEMs

A comparison of output from the baseline and cleaned SMT engines shows universal improvements for the AEMs – see Figure 4. At system level, the BLEU scores improved from .45 to .49 ($p > .05$); at segment levels both GTM and TER also show improvements from .47 to .50, and from .48 to .39 respectively ($p > .05$). These findings show that in addition to the above results concerning reductions in the prevalence of inconsistencies, the quality of the translations did not suffer. On the contrary, considerable improvements were made for all metrics, for which there were strong significant correlations, e.g. between GTM and TER where $r = -.85$, $p < .05$.

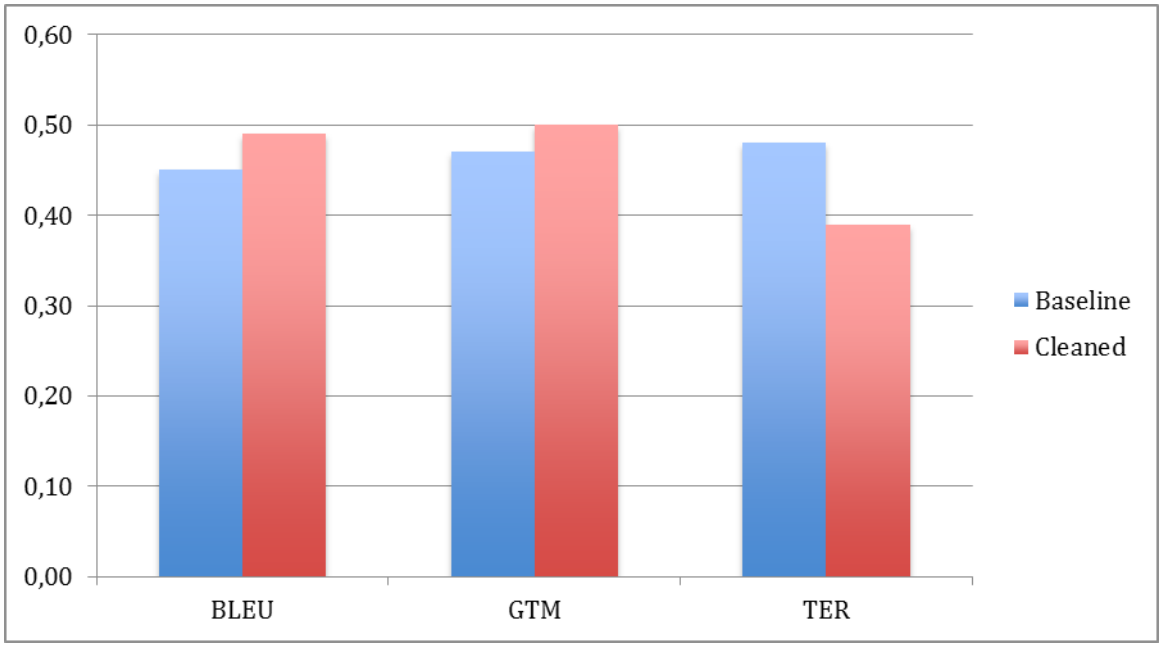

Figure 4: AEM Scores for Each Engine

Conclusion

Moorkens (2012) showed the prevalent categories of inconsistencies within contemporary commercial TM data and suggested that inconsistency introduced in translations of repeated ST segments would result in increased costs, growing year on year. Methods of improving TM consistency should thus have financial benefits. Given that TM data is also used to train SMT engines, the question arises as to whether more consistent TM data result in higher quality SMT output. This initial study has investigated whether using SMT output to improve the consistency of TM data could instigate a virtuous circle of improvement, where more consistent TM data can in turn result in better quality SMT output. The study has shown promising results when our methodology was applied to a single English-to-German TM. The results are mitigated somewhat by the propagation of some SMT errors in the TM data, and by potential difficulties in automating the process in cases where, for example, the SMT output has changed word order. Given current interest in automatically integrating TM and SMT workflows, however, we see our line of research as worthy of more attention. In continuing this research, we are currently applying this method to other unseen English-to-German and English-to-Japanese TMs in an attempt to improve external validity. Should these further tests show positive results, we believe that in research and commercial applications, this method could result in quantifiable qualitative improvements in translation using both TM and SMT, and, therefore, represent savings on resources such as translation time, cost, and effort.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. Thanks to Prof. Andy Way for the use of the SmartMATE system.

Bibliography

Debove, Antonia, Furlan, Sabrina, Depraetere, Ilse (2011), « A contrastive analysis of five automated QA tools (QA Distiller. 6.5.8, Xbench 2.8, ErrorSpy 5.0, SDLTrados 2007 QA Checker 2.0 and SDLX 2007 SP2 QA Check) » in Depraetere, I. (ed.), Perspectives on Translation Quality, Walter de Gruyter, pp161-192.

Doddington, George (2002) « Automatic Evaluation of Machine Translation Quality Using N-Gram CoOccurrence Statistics » in Proceedings of The Second International Conference on Human Language Technology, San Diego, CA, pp. 138-145.

Doherty, Stephen (2012) « Investigating the Effects of Controlled Language on the Reading and Comprehension of Machine Translated Texts » PhD Thesis. Dublin City University.

Doherty, Stephen, Kenny, Dorothy, & Way, Andy (2012) « Taking statistical machine translation to the student translator » in AMTA-2012: the Tenth Biennial Conference of the Association for Machine Translation in the Americas Proceedings, San Diego, CA, October 28 – November 1, 10pp.

Guerberof, Ana (2012) « Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation » PhD Dissertation. Universitat Rovira i Virgili, Tarragona.

He, Yifan (2011) « The integration of machine translation and translation memory » PhD thesis, Dublin City University.

Hearne, Mary & Way, Andy (2011) « Statistical Machine Translation: A Guide for Linguists and Translators » Language and Linguistics Compass 5, pp205-226

Koehn, Philipp (2010) « Statistical machine translation » Cambridge: Cambridge University Press.

Melamed, I. Dan, Green, Ryan & Turian, Joseph P. (2003) « Precision and Recall of Machine Translation » in Proceedings of HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series, Edmonton, Canada, pp. 61-63.

Moorkens, Joss (2012) « Measuring Consistency in Translation Memories » PhD Thesis. Dublin City University.

O'Brien, Sharon (2006) « Eye-tracking and Translation Memory Matches » Perspectives: Studies in Translatology, 14 (3), pp185-205

Ozdowska, Sylwia & Way, Andy (2009) « Optimal Bilingual Data for French–English PB-SMT » in Proceedings of the 13th Annual Conference of the European Association for Machine Translation, EAMT'09, Barcelona, Spain.

Papineni, Kilshore, Roukos, Salim, Ward, Todd, Zhu, Wei-Jing (2002) « Bleu: a method for automatic evaluation of machine translation » in 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp 311–318.

Rieche, Adriana Ceschin (2004) « Memória de tradução: auxílio ou empecilho? » MA Thesis. Pontifícia Universidade Católica do Rio de Janeiro.

Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea, Makhoul, John (2006) « A Study of Translation Edit Rate with Targeted Human Annotation » in Proceedings of 7th

Conference of the Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, pp. 223-231.

Turian, Joseph P., Shen, Luke & Melamed, I. Dan (2003) « Evaluation of Machine Translation and its Evaluation » in Proceedings of MT Summit IX, New Orleans, USA, pp. 386-393.

Way, Andy, Holden, Kenny, Ball, Lee, Wheeldon, Gavin (2011) « SmartMATE: Online Self-Serve Access to State-of-the-Art SMT » in Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators", JEC 2011, Luxembourg, pp43-52.