Machine Translation technology in advanced language teaching and translator training: a corpus-based approach to post-editing MT output

Bogdan BABYCH, Anne BUCKLEY, Richard HUGHES, Svitlana BABYCH
Centre for Translation Studies
University of Leeds
Leeds, UK
{b.babych, a.buckley, r.j.hughes, s.babych}@leeds.ac.uk

At advanced stages of language learning students are refining their skills in producing idiomatic and stylistically appropriate texts, with acceptable usage patterns, conventional terminology and a broad range of lexical resources, such as synonyms, collocations and complex lexico-grammatical constructions. Specifically, this level of proficiency is particularly important in translator training (c.f. Kuebler, 2011, Aston, 1999).

Creating or finding teaching materials and exercises for advanced learners is a challenging task for language tutors, because at this level traditional textbooks are not sufficient. Tutors usually rely on authentic texts and give students feedback on their productive skills. However, there are a number of practical and ethical issues which limit the use of this critical feedback: (a) it needs to be confidential: usually students can access only their own corrected work, which makes it difficult to discuss individual students' errors with the whole class; (b) students have limited possibilities to distance themselves from the text they produce and review the errors critically.

In our paper we explore techniques for developing non-native linguistic intuition via post-editing of imperfect Machine Translation (MT) output, using concordance, collocation and terminological searches in large-scale monolingual and bilingual corpora with the goal of arriving at a more fluent and coherent text. During the module we cover a range of text types and genres (journalistic, administrative, technical and literary) – specifically those that are discussed within other modules on Specialised Translation in our MA programme in Translation Studies. The texts, originally written in English, are automatically translated into another language, such as German or Russian, with an MT system (e.g. Systran) and then back-translated with MT into English. Alternatively we use a human translation of a text as input for MT, and use the text originally written in English as a reference.

Machine Translation technology was originally designed to generate rough translation for people who do not understand the source language, or to produce editable draft for bilingual professional translators (Hutchins and Somers, 1992). It has been suggested that there are similarities between errors made by MT systems and those made by non-native speakers (Lee et al, 2007). For our purposes it is important that MT output contains the original message, but with its fluency disrupted on the lexical, collocational or stylistic levels. The advantage of using MT in the classroom is that students can critically review the MT output, discuss potential solutions in a group with the tutor, and check their decisions by doing corpus-based research. The learning objective is to match the construction, lexicon, terminology and stylistic resources of the reference text or come up with alternative acceptable solutions, using our corpus-based tools built around the Corpus Workbench search

engine (Christ, 1994; Evert and Hardy, 2011). These tools include monolingual and bilingual concordances, terminological and collocation searches in large corpora (between 100 million and 1 billion words). Students explore usage patterns of specific linguistic expressions beyond the given text, try to find appropriate solutions to non-trivial editing problems, and then discuss their findings with the tutors and other native speakers. This type of work develops language skills at an advanced level via discovery learning and promoting learner autonomy. This gives the students more confidence in their ability to identify non-fluent phrases and find good near-native-speaker expressions for correcting these dysfluencies.

In the paper we explore the advantages of using post-editing of machine translation output combined with corpus-based translator training in the classroom for advanced language learners. We propose this new type of task for translator training and describe how the MT technology has to be adapted to better meet the requirements of this task, by being embedded into corpus-based resources and offering alternative translations for segments. For this purpose having non-idiomatic output is surprisingly an advantage, not a shortcoming of the MT system.

References

Aston, G. (1999). 'Corpus use and learning to translate.' *Textus* 12: 289-313.

Christ, O. (1994). 'A modular and flexible architecture for an integrated corpus query system.' Presentation at COMPLEX'94, Budapest, Hungary.

Evert, S. & Hardie A.. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the New Millennium.' Presentation at Corpus Linguistics 2011, University of Birmingham, UK.

Hutchins, W.J. & Somers, H. L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Kuebler, N. (2011). 'Working with corpora for translation teaching in a French-speaking setting.' In: Frankenberg-Garcia, A. Aston G. & Flowerdew L. (eds) *New Trends in Corpora and Language Learning*. London: Continuum: 62-79.

Lee, J., Zhou M., & Liu X. (2007). 'Detection of non-native sentences using machine-translated training data. *Proceedings of NAACL HLT 2007, Companion Volume*. 93-96.