Advanced learners' errors in correcting Machine Translation output: comparative corpus-based analysis

Bogdan Babych
University of Leeds
b.babych@leeds.ac.uk

Anne Buckley
University of Leeds
a.buckley@leeds.ac.uk

Svitlana Babych
University of Leeds
s.babych@leeds.ac.uk

1 Introduction

This paper describes an experiment on representing, annotating and analysing errors made by language learners who correct the output of Machine Translation (MT) systems. In our previous work (Babych et al., 2012) we presented the method of using error correction in advanced stages of language learning and translation training, where negative linguistic evidence is automatically generated by rule-based MT systems. MT output usually contains the original message, but with its fluency disrupted on the lexical, collocational or stylistic levels. By correcting MT errors the students are refining their skills in producing idiomatic and stylistically appropriate texts, with acceptable usage patterns, terminology, synonyms, collocations and lexico-grammatical constructions appropriate for the situation and linguistic context. This high proficiency level is particularly important for trainee translators (Kuebler, 2011; Aston, 1999). Our students critically review the MT output, discuss potential solutions in a group and/or with the tutor, and check their decisions by doing corpus-based research. In our method MT is used not simply as a useful dictionary alternative, but for systematically generating negative linguistic evidence (cf. Landure & Boulton 2010). Even though using ill-formed L2 can be counterproductive in the initial stages of foreign language learning (Somers, 2004), in the advanced stages negative linguistic evidence is useful, since students are aware of contrastive differences between languages and consciously take control over developing their productive skills in autonomous learning.

In this paper we describe the corpus format we use to represent MT output errors, which are categorized and aligned with corresponding students' (successful and unsuccessful) error corrections, and also classified and compared to the initial set of MT errors. Further we present an analysis of different error types in the corpus. The results inform the way in which we apply the proposed method in our 1-semester MA module English for Translators taught for Translation Studies students at the University of Leeds.

2 Error representation format and categorisation scheme

Students receive MT-generated texts as homework and do corrections in 3 or 4 groups on Wikis on the VLE. For the following class their corrections are annotated with the following colour coding: Green - excellent solution; White - acceptable solution; Yellow- could be improved (e.g. meaning correct but not very idiomatic); Red - wrong solution.

For the corpus we align initial MT errors with students' correction solutions submitted by each of the groups, as shown in Figure 1.

MT Output			
He led	me	**strongly at**	Jon, **who speak other** rich-looking couple.
TENSE		COLLOC	SYNTAX
Group1			
			speaking to another rich-looking couple.
	me directly to Jon, who		
He led		was	
Group2			
			Jon, who is talking with a rich-looking couple.
	me		
		strongly toward	
He pulled			
Group3			
He takes me			to another rich-looking couple.
		aggressively to Jon speaking	

Figure 1: Alignment of MT errors and solutions

We annotate students' and MT errors using an error categorization scheme inspired by (James, 1998). The scheme is based on linguistic levels of errors (morphological, syntactic, lexical), but also takes into account the frequency of different error types. For example, collocation errors are a type of lexical error, but we annotate them separately because this category is very frequent. Table 1

shows examples of annotated error types.

Error type	Example
COLLOC	***Exaggerated*** *lipstick (=too much lipstick)*
LEXICAL	*Put hand on her shoulder with a* ***possessiveness*** *(=possessively)*
PREPOS	*Ten yards* ***at*** *my left (=on my left)*
TENSE	*Before I can stop him he* ***led*** *me to Jon (=leads)*
SYNTAX	*I* ***have not to him spoken*** *yet (=I have not spoken to him)*
ARTICLE	*You can make* ***better decision*** *(=a better decision)*
MORPH	*We are open for* ***negotiation*** *(=negotiations)*

Table 1: Examples of error types

3 Corpus-based error analysis

When working in groups, students can either miss, or successfully identify but incorrectly change, or finally – successfully correct MT errors. Table 2 shows percentages of such cases for each of the error types in MT output.

	not found	*changed (wrong)*	*corrected*	*Total*
COLLOC	18.9%	39.2%	41.9%	44.0%
LEXICAL	43.5%	30.4%	26.1%	13.7%
PREPOS	20.0%	33.3%	46.7%	8.9%
TENSE	50.0%	8.3%	41.7%	7.1%
SYNTAX	0.0%	52.0%	47.7%	26.2%
Total	19.60%	38.7%	41.7%	100.0%

Table 2: Students' initial correction of MT errors

It can be seen from the table that while students always identify syntax errors, in around 20% of cases they do not see that there is a problem with a collocation or a preposition; when MT errors are correctly identified, about 50% of students' initial changes are correct.

Finally, we annotated error relation patterns in MT and student texts. We recorded which types of MT errors resulted in which types of the student errors, and which errors were corrected or emerged from the correct structures ('0' symbol in Table 3 shows absence of errors):

Pattern	Percentage
COLLOC>COLLOC	17.3%
COLLOC>0	10.9%
SYNTAX>0	10.9%
LEXICAL>LEXICAL	10.0%
0>COLLOC	5.5%
LEXICAL>0	5.5%
0>TENSE	4.5%
TENSE>TENSE	4.5%
PREPOS>PREPOS	3.6%
SYNTAX>COLLOC	3.6%
0>PREPOS	2.7%
PREPOS>0	2.7%

Table 3: Frequent error relation patterns (MT>Students)

It can be seen from Table 4 that the most frequent patterns are COLLOC>COLLOC – non-identified or wrongly corrected collocation error, followed by corrected collocation and syntax errors; in 5.5% of cases a new collocation error was introduced. This highlights the fact that collocations remain one of the most serious challenges for advanced language learning. Other patterns for newly introduced student errors are 0>TENSE and 0>PREPOS, which shows that these types of problems also have high priority for advanced learners.

References

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

Babych, B., Buckley, A., Hughes, R & Babych, S. 2012. Machine Translation technology in advanced language teaching and translator training: a corpus-based approach to post-editing MT output. In: Proceedings of of TALC 2012 : Teaching and Language Corpora Conference. Warsaw, Poland on 12th - 14th July 2012.

James, K. 1998. *Errors in language learning and use. Exploring error analysis*. London and New York: Longman.

Kübler, N. 2011. Working with corpora for translation teaching in a French-speaking setting. In *New trends in corpora and language learning*,A. Frankenberg-Garcia, G. Aston & L. Flowerdew (eds.), 62-79. London: Continuum.

Landure, C., & Boulton, A. 2010. Corpus et autocorrection pour l'apprentissage des langues. *Asp* 57: 11-30.

Somers, H. 2004. Does machine translation have a role in language learning? In *Proceedings of UNTELE 2004: L'Autonomie de l'Enseignant et de l'Apprenant face aux Technologies de l'Information et de la Communication – Teacher and Learner Autonomy vis-a-vis Information Communication Technology*, Compiègne, France, 28.