Translation Quality Evaluation of MWE from French into English using an SMT system

Emmanuelle Esperança-Rodier
Université Grenoble Alpes – LIG/GETALP
Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr

Johan Didier
Université Grenoble
Johan.Didier@univ-grenoble-alpes.fr

Abstract

Nowadays, Statistical Machine Translation (SMT) is widely available. Nevertheless, using Machine Translation at its best is not an easy task. Structures appearing sporadically trigger most of the regular mistakes of SMT systems. We work on one of those structures: the MultiWord Expressions (MWE). Our study aims at evaluating the quality of MWE translation obtained using SMT.

Firstly, we present the process of our quality evaluation of the English translation got via an SMT system created using Moses Toolkit (Koehn et al., 2007), of one French technical document. On the French document, MWE have been semi-automatically annotated according to their type (Tutin et al., 2015). Secondly, we describe the linguistic criteria of Vilar's classification of translation errors (Vilar et al. 2006) as well as the adaptation we had to perform to use Blast (Stymne, 2011). Thirdly, we analyse the global results of our quality evaluation before going into details, in our fourth part, on one particular type of MWE, which is the Full Phraseme one. We finally show that most of the French MWE are translated into English MWE, and that we need to implement in further work a collaborative error annotation tool.

1 Introduction

Machine Translation (MT) nowadays is widely available for a large set of people; professional translators, students, researchers, common people. Nevertheless being able of using MT at its best is not an easy task. Most of the MT systems make regular mistakes, which means that there are typical syntactic structures, lexical items they cannot deal properly with. Therefore, the end user has to be able to detect and to correct those mistakes. The ability of the user to detect and correct the mistakes relies on his/her language skills, which means that some people would have the necessary skill to detect that the sentence is not grammatically correct and would correct it, but some others would not know how to correct it. The language skill level related to the post edition of MT is an interesting topic, which we won't discuss in this paper for a matter of length.

If we go back to the mistakes themselves, some of them are due to the fact that as MT systems use probabilistic algorithm they cannot focus on structures that does not appear very often. In this article, we are going to take this problem into account and thus deal with one of those structures that are the MultiWord Expressions (MWE). MWE are very common but not always in proportions that are sufficient for MT to give successful translations. However, as MWE are typical of the language, if they are mistranslated, the end user would consider the whole translation as a poor one even if it is not the case. Among others, previous work from Ramisch et al. (2013) on one type of MWE has already depicted the complexity of MWE translation.

We have to keep in mind that, in this work, only the quality of the MWE translation has been addressed. Consequently, the information of a sentence not being translated correctly is not given.

Proceedings of the 38th Conference Translating and the Computer, pages 33–41,
London, UK, November 17-18, 2016.

In order to study the quality of MWE translation, we based our work on the analysis of a corpus.

Our corpus is made of one technical document (12,566 words) written in French on which MWE have been semi-automatically annotated according to their type. As we based our work on a linguistic approach, we used a script to locate the more obvious MWEs, which we validated or not, and then we manually annotated those which were not found automatically or for which the type was not so clear-cut. This way we were able to refine the MWE type definitions as well as work on the inter annotator agreement. We have thus decided on purpose not to use automatic tools such as Ramisch et al. (2010) proposed.

2 Methods

2.1 MWE annotation

We semi-automatically annotated, as we said, the MWE present in our corpus according to nine types described in (Tutin et al., 2015). In this previous paper, MWE were addressed as "multiword elements that includes several graphical units, separated by blanks or hyphens, or separated by several other words not included within the MWE".

Out of those nine types, we decided to focus on five of them, which are the following ones:

- Function words (F),
- Full Phrasemes (PH),
- Collocations (C),
- Technical Terms (T) and
- Named Entities (EN).

Tutin (2015) defined those five types in a draft of Annotation guidelines for multi-word expressions. In the interests of brevity, we will just give a rough definition for each of the five types here.

First, Function words are characterized by a vague, and mainly functional, meaning. They include grammatical words such as conjunctions e.g. even if, or among others, prepositions e.g. in front of.

Second, Full Phrasemes include MWEs which are not compositional, e.g. couch potato, and/or are words, mainly nouns, which refer to specific referent, e.g. death penalty.

Collocations include frequent compositional expressions, e.g. heavy smoker.

Technical Terms, a subtype of full phrasemes, are mainly nominal full phrasemes typical of specialized corpora.

On top of the type, the MWE annotation also consists of the part of speech of the MWE, the part of speech of the elements of the MWE, and the overlapping of MWE is also annotated…

The French annotated document has been translated into English by a MT system created using Moses Toolkit (Koehn et al., 2007).

Furthermore, we have decided that the quality evaluation of the obtained translation, would not be done via automatic metrics but using more linguistics criteria such as those defined in (Vilar et al., 2005), which we are going to describe in the following section.

2.2 Error type classification

Actually, the criteria used in Vilar's classification of translation errors, as described in Figure 1, suit pretty well the linguistic evaluation we want to perform.

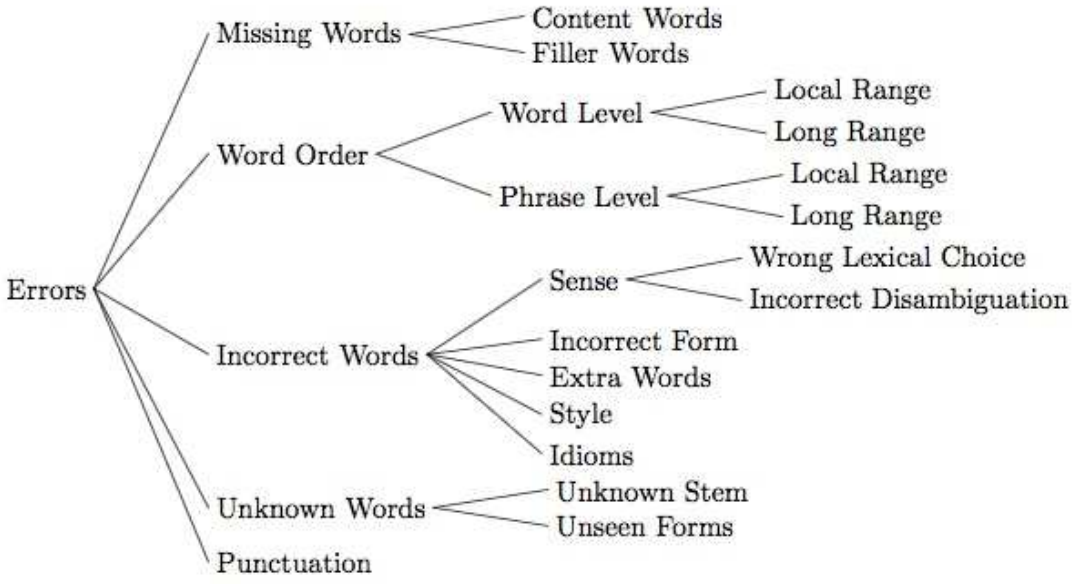

Figure 1. Classification of translation errors.

Five main translation error categories have been identified, namely:

- Missing Words, for words which have not been translated,
- Word Order, for a wrong order of the words in the translated sequence,
- Incorrect Words, for a mistranslation,
- Unknown Words, for words that were not known by the system and thus left in the source language,
- Punctuation, when punctuation rules of the target language were not respected.

As regards to the first two translation-error categories, Missing Words and Word Order, subcategories have been created to refine the error class. For the Missing Words error category, the distinction between Content and Filler words allows to see whether the missing word was meaningful or not. This subcategory illustrates the fact that the full meaning of the sentence was kept or not, which is obviously the aim of a quality translation evaluation. As regards to the Word Order error category, the Word or Phrase level subcategory shows if the translation error entails a reordering of the words themselves, or a reordering of phrases. It is well addressed to SMT evaluation as it permits to locate at which level the system failed, lexical level or syntactic level.

Looking at the third translation error category, Incorrect words, we can see that there are several subcategories aiming at distinguishing the reason of the mistranslation, which can be due to the fact that the system was not able to disambiguate properly the meaning of a source word nor to produce the right form of the word, although the base form of the word was well translated.

For the fourth translation error category, Unknown words, we can distinguish whether the stem of the words was known by the system or not.

And finally, the fifth translation error category, which is Punctuation, did not receive full attention from our part.

In addition, the part of speech of the word linked to the translation error category or subcategories, is also given. Thanks to Vilar's classification we were able to undergo a rich translation error annotation.

However, we wanted to track some more pieces of information. We have therefore added to the Vilar's classification the four following features. Hereafter, we will describe those additional features by defining them and giving their annotation abbreviation.

Firstly, we added the type of the MWE. The annotation of MWE in the source document was already giving that information, but due to technical difficulties we were not able to recover that information after translation. This is why we have decided to integrate the type of MWE into the translation error categories at the first level that is to say prior to any translation error category. We have used the abbreviation given in paragraph 2.1, e.g. EN for Named Entities…

Secondly, as we wanted to focus tightly to the translation quality evaluation of MWE, we wanted to refer to in-use translations of MWE. Consequently, for the most frequent MWE found in our source corpus, we have extracted from the bilingual concordancer Tradooit (https://www.tradooit.com), the related translations, considering them as the attested translations. We have thus integrated a category showing that the translation of the MWE was an attested one or not. We used the TA abbreviation when the translation was attested and the TNA abbreviation when it was not an attested translation.

Thirdly, we addressed the translation quality criteria by distinguishing four quality levels. When the source MWE was well translated, we used the BT abbreviation. When the source MWE was wrongly translated then we used the abbreviation MT for wrong translation. Then when the translation had to be edited but the meaning of the source sentence was kept, we used the abbreviation RevPres. And when the translation had to be edited but the meaning of the source sentence was not kept, we used the RevNPres abbreviation.

Finally, we wanted to annotate the fact that the translation was also a MWE in the target language, or not. We respectively used the abbreviations MULT and NONMULT. Consequently, Vilar's classification of translation errors has been extended to this scheme, Figure 2:

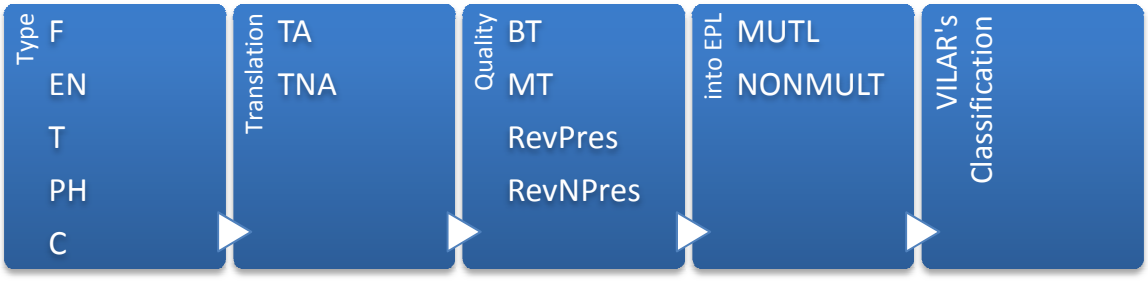

Figure 2. Extended Classification

2.3 Evaluation tool

The English translation has then been evaluated using BLAST (Stymne, 2011), an open source tool for the annotation of translation errors. We have chosen this tool as it has already been used for former MWE translation evaluation; and that it uses as a standard, and among others, the Vilar's Classification of translation errors, even if it can be used with any other hierarchical error classifications. Blast is also highly adapted to any evaluation purposes as it is not linked to the information provided by any specific MT. Furthermore, Blast is easy to

use because of its graphical interface. Among all the above, we have seen in Blast the way to add new annotations, edit existing annotations and also to search among the annotations.

Nevertheless we have experienced mainly two kinds of problems with the evaluation criteria. Firstly, we have encountered the borderline case of annotating several MWE in a same sentence. When several MWEs were present in a same sentence, it was impossible for us to link the translation error category respectively to the related MWE.

Secondly, as Blast ignores identical annotation, it seems impossible to have access to several error types at the same time for the same MWE. Also because of the hierarchical structure of the tool, we had to declare all the possible path of translation error annotation, thus having a linear path.

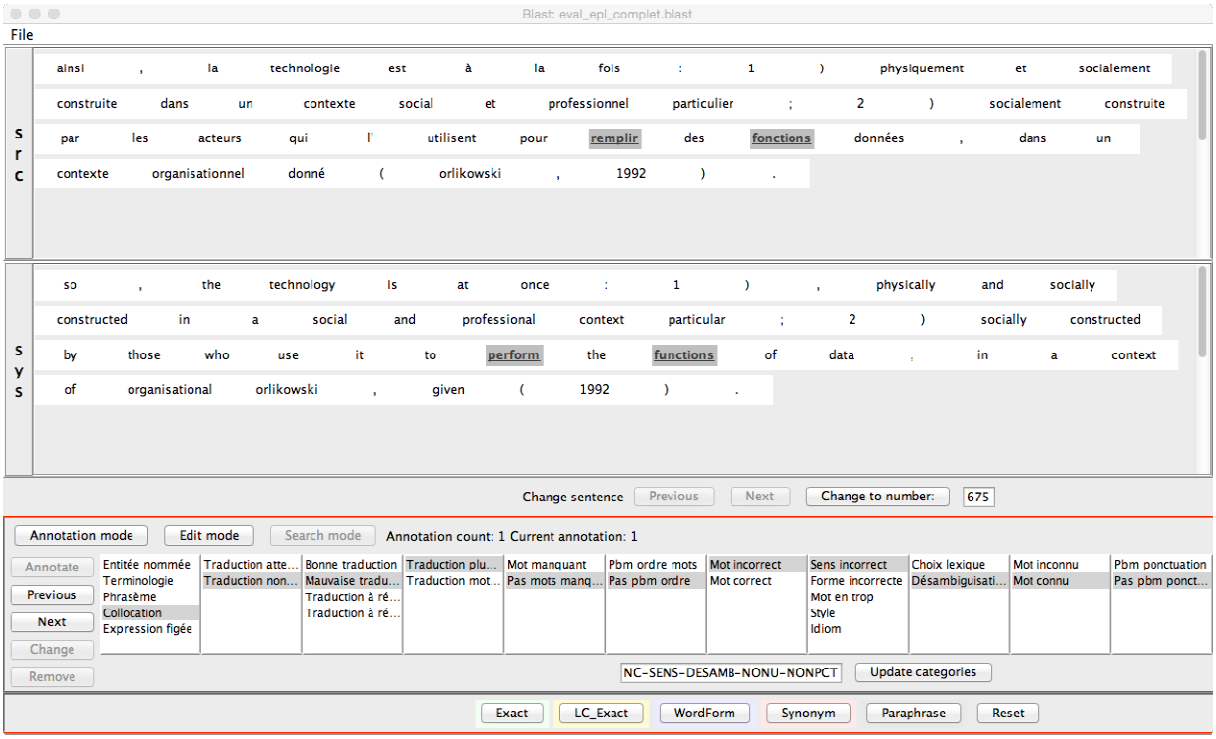

Figure 3. Blast Screenshot

The translation error annotation has thus been difficult. We could also mention that as the annotator has to make a decision, he cannot use the tool to trigger the attention on a difficult MWE to annotate thus asking for some help from another annotator. The possibility of having a collaborative tool would have been greatly appreciated for the MWE quality translation evaluation.

3 Results

3.1 Global results

Once the English translation evaluated, we have been looking at the global results which we have collected in Table 1 and Table 2 below.

As previously mentioned, out of the nine types of MWE, we have only worked on five of them. Those five types of MWE are represented thanks to their abbreviation, see above section 2.1, in the first column of Figure 4.

MWE type	Total number of occurrences	Attested Translation *TA*	Good Translation *BT*	Good Translation/ Attested Translation *BT/TA*	Good Translation/Non Attested Translation *BT/TNA*	Bad Translation/ Attested Translation *MT/TA*	Bad Translation/ Non Attested Translation *MT/TNA*
PH	135	64.4	64.4	96.6	6.3	0	66.7
C	308	63.3	72.1	96.9	29.2	0	22.1
F	202	78.2	81.2	98.1	20.5	0	36.4
EN	9	77.8	66.7	85.7	0.0	0	100.0
T	51	58.8	64.7	100.0	14.3	0	28.6

Table 1: Global results for 5 MWE types on Good Translation

Back to Table 1, if we look at the total number of occurrences per MWE type, it appears that we have mainly found in our corpus the Collocation type (C), with 308 occurrences. While the type of MWE we have found in the smallest quantity is the Named Entities (EN), 9 occurrences. Nonetheless, the ratio of bad translation over the not attested translation (MT/TNA), given in the eighth column, for EN is the highest with 100%. It means that when a MWE of the Named Entity type has been translated by a non attested translation it was also always a bad translation.

On the contrary, when a MWE of the Technical Term Type (T) was translated by an attested translation, it was always a good translation. The ratio good translation over attested translation, BT/TA, is 100% as mentioned in column number 5.

As a validity control, when an attested translation was given, it has never been a bad translation. The ratio MT/TA, appearing in the seventh column, is 0% for any type of MWE.

MWE type	Total number of occurrences	Attested Translation *TA*	Good Translation *BT*	Translation to be edited but source meaning kept/ Attested Translation *RevPres/TA*	Translation to be edited but source meaning kept /Non Attested Translation *RevPres/TNA*	Translation to be edited but source meaning not kept/Attested Translation *RevNPres/TA*	Translation to be edited but source meaning not kept/Non Attested Translation *RevNpres/TNA*
PH	135	64.4	64.4	1.1	18.8	0	8.3
C	308	63.3	72.1	1.5	32.7	0	13.3
F	202	78.2	81.2	1.3	40.9	0	2.3
EN	9	77.8	66.7	0.0	0.0	0	0.0
T	51	58.8	64.7	0.0	47.6	0	9.5

Table 2: Global results for 5 MWE types on Translation to be edited

In the same way, for validity control, we can notice on Table 2 that for an attested translation given, no translation evaluated to be edited with the meaning not being kept were found. For all the MWE types, the ratio RevNPres/TA, given in the seventh column, equals 0.

4 Detailed results focusing on Full phrasemes

We have decided to focus on the Full Phraseme type as they are the most difficult MWE to understand for non-native speakers. In the following, we are going to cover specific columns of Table 1 and Table 2, explaining the results by giving five examples, made of the French source, and of the English SMT Translation, and also of the related error annotation path for each example.

Actually, referring to Table 1, we find that 64%, see third column, of the Full Phrasemes were correctly translated, as we can see in the example 1. We can notice in that first example that the whole translated sentence is not correctly translated. As we already said, we have only processed to the quality evaluation of the MWEs, not to the evaluation of the whole sentence.

Example 1
French: […] leur statut d'embauche plus précaire fait en sorte qu'ils sont soumis à une forte pression […]
English SMT Translation: […] their status of employment more precarious done in such a way that they are under pressure […]
Error Annotation path: ph-TA-BT-MULT

In the fifth column of Table 1, we can see that more than 96% of Full Phraseme MWEs have been well translated when an attested translation existed. This result corresponds to the trend of the whole study. Nevertheless, roughly 6%, as written in the sixth column, of the Full Phraseme MWEs were well translated while the translation was not one of the attested ones. If we look at the example 2 below, the Full Phraseme faire état has not been translated in one of its attested translation but even so, it has been well translated by to present.

Example 2
French: […] nous allons ensuite faire état des méthodes […]
English SMT Translation: […] then we are going to present the methods […]
Error Annotation path: ph-TNA-BT-MULT

Going further in Table 2, we notice in the fifth column, that 1% of the translations has been evaluated as needing to be edited while the meaning was kept when an attested translation existed. Actually, example 3 shows that the Full Phraseme mis en oeuvre has been correctly translated as regards to its attested translation and meaning, but that its form was incorrect as the passive voice to be implemented was used in the translation while it should not.

Example 3
French: […] et les principales adaptations requises mises en oeuvre […]
English SMT Translation: […] and the major adaptations required to be implemented […]
Error Annotation path: ph-TA-REV_PRES-MULT-NONM-NONO-INC-FORME-NONU-NONPCT

Going to the next column, it emerges that a bit less that 19% of the translations have been evaluated as needing to be edited while the meaning was kept when the translation was not an attested one. As shown in Example 4, the Full Phraseme MWE pris en charge has not been translated by one of its attested translation, and that the translation proposed, i.e. taken over has taken the wrong lexical choice.

Example 4
French: […] la mécanisation et l' automatisation des procédés de travail dans l' industrie manufacturière ont été prises en charge par la production à la chaîne […]
English SMT Translation: […] the mechanisation and automation of working processes in the manufacturing industry have been taken over by the production of the chain […]
Error Annotation path: ph-TNA-REV_PRES-MULT-NONM-NONO-INC-SENS-LEX-NONU-NONPCT

In the last example, and referring to the last column of Table 2, only 8% of the translations have to be edited when it was not an attested translation of the Full Phraseme MWE. The Full Phraseme MWE en bout de ligne has been translated into at the end of the line and thus evaluated as needing to be edited with the meaning not being respected because of the use of Incorrect Words due to an Incorrect Disambiguation related to the Sense subcategory.

Example 5
French: mais en bout de ligne […]
English SMT Translation: but at the end of the line […]
Error Annotation path: ph-TNA-REV_NONPRES-MULT-NONM-NONO-INC-SENS-DESAMB-NONU-NONPCT

It would have been useful to check if the same MWE appeared several time in the document to verify the translation consistency. Is the MWE always translated in the same way, according to its place in the sentence, or to the syntactic pattern in which the MWE has been found? As Blast only allows to search for error categories, and that it does not permit to search for words or patterns, we could not proceed to such an investigation.

5 Conclusion

As a first conclusion, we have found that 80% of the MWE found in the French text were translated into MWE in English. As regards to the studied MWE types, the good translation rate is acceptable, showing that work has to be done in order to improve it.
As our corpus is not really big, one text of roughly 12,500 words, we want to draw the reader attention to the fact that this work is a first investigation of the MWE translation quality. Also, for some translation error annotations, the error annotation path was really long and thus some inconsistencies could arise.

Our second conclusion is then, that we would need another tool specifically dedicated to translation error annotations, with the possibility of selecting the source text and its target translation and to assign a translation error type. It will help identifying patterns in which specific translation error categories occur more often.

An extended work will thus consist in specifying a new collaborative tool dedicated to translation error annotation. Finally, a further work will be dedicated to deeply look at the quality translation results of the different MWE types studied.

Acknowledgements

This work has benefited from the AIM-WEST project (http://aim-west.imag.fr), which deals with the analysis and integration of MultiWord Expressions (MWEs) in speech and translation.

References

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi,Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst, 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions , pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Carlos Ramisch, Aline Villavicencio, Christian Boitet, 2010. mwetoolkit: a Framework for Multiword Expression Identification Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta

Carlos Ramisch, Laurent Besacier, Alexander Kobzar, 2013. "How hard is it to automatically translate phrasal verbs from English to French?", MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, France, September 2013.

Sara Stymne, 2011. Blast: A tool for error analysis of ma- chine translation output. In Proc. of the ACL 2011 System Demonstrations, pages 56–61, Portland, OR, USA, Jun. ACL.

Agnès Tutin, Emmanuelle Esperança-Rodier, Manuel Iborra, Justine Reverdy, 2015, Annotation of multiword expressions in French. Malaga, Espagne, Actes de la conférence Europhras2015, Juin 2015.

David Vilar, Jia Xu, Luis Fernando D'Haro et al., 2006. Error analysis of statistical machine translation output. In : Proceedings of LREC. 2006. p. 697-702.