Multi-word expressions in user-generated content: How many and how well translated? Evidence from a post-editing experiment

Seretan, Violeta

Multi-Word Expressions in User-Generated Content: How Many and How Well Translated? Evidence from a Post-editing Experiment

Violeta Seretan
Department of Translation Technology
Faculty of Translation and Interpreting, University of Geneva
40 Bvd. du Pont-d'Arve, CH-1211 Geneva, Switzerland
Violeta.Seretan@unige.ch

Abstract

According to theoretical claims, multi-word expressions are pervasive in all genres and domains, and, because of their idiosyncratic nature, they are particularly prone to automatic translation errors. We tested these claims empirically in the user-generated content domain and found that, while multi-word expressions are indeed common in this domain, their automatic translation is actually often correct, and only a modest amount – about one fifth – of the post-editing effort is devoted to fixing their translation. We also found that the upperbound for the increase in translation quality expected from perfectly handling multi-word expressions is 9 BLEU points, much higher than what is currently achieved. These results suggest that the translation of multi-word expressions is nowadays largely correct, but there is still a way to go towards their perfect translation.

1 Introduction

The literature on multi-word expressions (henceforth, MWEs) abounds with claims on the pervasiveness of such expressions in language, as well as on the difficulty of translating these expressions automatically. It is held that multi-lexeme units are of the same number of magnitude as single-lexeme units (Jackendoff, 1997), or even one order of magnitude more numerous (Mel'čuk, 1998). Also, it has been suggested that no single utterance is totally free of MWEs (Lea and Runcie, 2002) and that MWEs are a major obstacle for achieving correct machine translation (Sag et al., 2002).

In order to cope with the MWE translation problem, the solutions adopted in the literature were to gather MWEs in the lexica of rule-based translation systems, as in Orliac and Dillinger (2003), or to find means to integrate them into the statistical machine translation (SMT) pipeline. In the latter case, the MWE integration is achieved either in the pre-editing stage in a so-called *words-with-spaces* approach (Carpuat and Diab, 2010), or in the training stage as supplementary "sentences" (Bouamor et al., 2012), or, again, by adding MWEs to the phrase table together with new features that constrain the SMT decoder to apply a MWE-compatible segmentation over a different one (Carpuat and Diab, 2010).

These integration strategies were reported to yield positive results. For instance, Orliac and Dillinger report "significant improvement in readability and perceived quality of the translation produced" (Orliac and Dillinger, 2003, 293). Also, a number of authors (see Table 1) report significant improvements of SMT quality in terms of BLEU score. The increase in translation quality which is due to MWE integration remains, however, quite limited with respect to the whole sentence score. As can be seen in Table 1, the increase is often less than one BLEU point. This modest increase seem to contradict the original statements on the pervasiveness of MWEs and their importance for achieving better translations.

The aim of our study is to look more in detail at the issue of MWE translation, in order to better understand the reasons beyond this positive but limited impact observed, and beyond the apparent contradiction with theoretical claims. First, we wanted to check whether MWEs are as pervasive in our data domain (the user-generated content domain) as the literature claims. Second, we wanted to investigate how many of the MWEs are badly translated. Third, we wanted to to see how much we could gain in terms of translation quality score if we had a perfect, 'oracle' translation for all MWEs in our test set (i.e., to determine the upperbound that a system could achieve, compared to the state of the art of one BLEU point).

Work	Language Pair	Test Set	Impact (BLEU Points)
Bai et al. (2009)	cn–en	NIST MT-06	0.22 (21.79 – 21.57)
Carpuat and Diab (2010)	en–ar	NIST MT-08	0.78 (31.27 – 30.49)
Liu et al. (2010)	cn–en	NIST MT-04	0.85 (30.47 – 29.62)
Tsvetkov and Wintner (2010)	he–en	not reported	0.10 (13.79 – 13.69)
Liu et al. (2011)	cn–en	NIST MT-04	1.10 (29.87 – 28.77)
	cn–en	NIST MT-2008	1.41 (19.83 – 18.42)
Bouamor et al. (2012)	fr–en	Europarl	0.30 (25.94 – 25.64)
Kordoni and Simova (2014)	en–bg	SeTimes	0.20 (23.9 – 23.7)

Table 1: Impact of multi-word expression integration on translation quality.

Our hypotheses, which were grounded on theoretical research, were the following:

1. MWEs are pervasive in user-generated content;
2. Most MWEs are badly translated;
3. Because of the above, post-editors spend a lot of effort correcting MWE translation errors.

We conducted empirical investigation on post-editing data available from the ACCEPT European project devoted to improving the translatability of user-generated content.[1] This data domain is less explored by existing MWE research, despite the fact that it represents one of the biggest challenges for natural language processing for the years to come. Our study provides evidence for the prevalence of MWEs in the social media genre represented by forum posts. While the pervasiveness assumption was confirmed, the other assumptions were challenged.

In the following sections, we describe the data (Section 2) and the investigation these data allowed, referring to above-mentioned hypotheses and to the findings obtained (Section 3). In the last section, we provide concluding remarks and ideas for future work (Section 4).

2 Data

The data used in our study is taken from a larger dataset available from the ACCEPT European projet (2012–2014) devoted to improving SMT of user-generated content. The dataset consists of 1000 technical forum posts in French, which have been automatically translated into English using a domain-adapted phrase-based statistical machine translation system (D41, 2013). The MT output has been manually corrected by a post-editor, a native speaker of English, paid for the task. The forum posts originate from the French chapter of the Norton Community Forum related to computer security issues.[2]

From the total 4666 corresponding translation segments, we randomly sampled 500 segments for the purpose of this study. Three of these segments turned out to be in English, as they were quotes of error messages included by forum users in their posts. After discarding these segments, we ended up with 497 segments in our test set. They total 5025 words and their length varies between one word (e.g., *Bonjour*, 'Hello') and 73 words, with an average size of 10.1 words/segment.

The example below shows a segment, its automatic translation, and the version corrected by the post-editor:

(1) a. *Source: Laissez tomber* ..... depuis 5 mois ..... j 'ai *résolu* la *question* hier
b. *MT: Let down* ..... for 5 months ..... I've *resolved* the *issue* yesterday
c. *Post-edited: Drop it* ..... after 5 months ..... I *fixed* the *issue* yesterday.

This example illustrates the discussion in Section 1. There are two MWEs in this segment, shown in italics. Both are badly translated and, as can be seen, their correction represent a large share of the total amount of corrections made by the post-editor.

3 Experiments and Results

In order to test the validity of the hypotheses put forward in Section 1, we conducted a series of experiments, summarised below.

[1] `www.accept-project.eu`. Accessed July, 2015.

[2] `http://fr.community.norton.com`. Accessed July, 2015.

3.1 Checking MWE pervasiveness

The assumption that MWE are pervasive in language is often taken for granted in the literature, as it seems superfluous to demonstrate the obvious. There are, however, studies in which authors provide empirical evidence supporting this claim. For instance, Howarth and Nesi (1996) came to the conclusion that "most sentences contain at least one collocation" (Pearce, 2001). While this holds for the general domain, little is known about the validity of this claim for the user-generated content domain.

In order to check the pervasiveness of MWEs in this domain, we proceeded to the manual identification of all MWEs in our test set of 497 segments. There are tools for MWE identification in text, or dictionaries we could have used to recognise MWEs in text. But we have chosen to annotate MWEs manually, mainly for the reason that user-generated content exhibit peculiarities which hinder the application of automatic methods: presence of slang, abbreviations, colloquial speech, errors at various levels (punctuation, casing, spelling, syntax, style, etc). Users of technical forums like the Norton Community forum are most likely to write in a hurry, because they are concerned by their problem at hand – for instance, by the fact that their computer crashes all the time, or the fact that some product they installed keeps on debiting their card monthly despite the subscription being cancelled. Their real concern is getting a solution to their problem as soon as possible, not the quality of their message. Any deviation from the norm is acceptable, as long as the message is understood by the community members.

Therefore, we chose to do the MWE annotation entirely manually, this way ensuring the accuracy of the annotation. The criterion used in deciding weather a combination is a MWE is the lexicographic criterion, i.e., we annotated a combination as MWE *iff* it was deemed worth of inclusion in a lexicon (in other words, it was not a regular combination). Despite this simple criterion, there is always some amount of uncertainty and subjectivity, as it is a well-known fact that MWEs are on a continuum from completely regular to completely idiosyncratic, and it is impossible to draw a clear-cut line between regular combinations and combinations which are MWEs (McKeown and Radev, 2000). In future studies, we may want to rely on judgements from multiple annotators in order to reduce the amount of uncertainty and subjectivity.

A specific annotation choice was necessary in the case of nested MWEs, i.e., when a MWE participates in another MWE. An exemple is provided below, in which *mise à jour* (lit., put to day, 'update') further combines with the verb *faire* to form a longer MWE, *faire mise à jour* (lit., to do update, 'to update'):

(2) Malgré les mises à jour faites (Démarrer>Windows Update), windows demande toujours les mêmes 2 maj

In this case, the decision taken was to count each MWE instance separately. Therefore, in this examples, we counted two MWEs.

Another specific annotation choice concerned the annotation of MWE reduced to abbreviations. In Example (2) above, there is a second instance of the MWE *mise à jour* occurring at the end of the sentence, as the abbreviation *maj*. In the framework of the ACCEPT project from which the data derives, abbreviations were treated as non-standard lexical items that have to be normalised in order to facilitate translation. As can be seen in Example (3), the post-editor understood the French abbreviation and corrected the MT output by proposing the full form equivalent in English, *update*. Influenced by the pre-editing approach adopted in the context of the project (*maj* → *mise à jour*), we decide to count abbreviations of MWEs as actual MWE instances.[3]

(3) a. *MT:* Despite updates made (Start > Windows Update), windows always ask the same 2 Shift

b. *Post-editor:* Despite the updates done (Start > Windows Update), windows always asks for the same two updates.

Given the methodological choices explained above, the statistics for the test set are as follows. The total number of MWEs in the 497 segments is 223. A number of 152 segments contain MWEs, which gives an average of 1.5 MWEs/segment. This might seem in line with known results from literature; however, reported to the total test size, the average is 0.4, lower than stipulated by litera-

[3] As an alternative, we could have ignored abbreviations, as one workshop attendee suggested. We maintain, however, that in a translation perspective, contracted MWEs require a full form version in order to facilitate their treatment.

ture. Since many segments are very short, we ignored segments that contained less than 100 characters, and got an average of 1.3 MWEs/segments for the remaining 91 segments.

Our results indicate that in the user-generated content domain, there seem to be less MWEs than in the general domain. However, the words participating in MWEs make up as much as 10.5% of the total words in the test set. Previous results for the general domain reported that MWEs account for just only 5% of the data in the NIST-MT06 test set (Bai et al., 2009). While a straight comparison is not possible because of the different methodologies used to recognise MWEs, the relatively high percentage obtained for the user-generated content domain suggests that MWE account for a larger portion of the data. From a translation perspective, it is important to focus on this portion of the data because it is likely to be more important in terms of comprehensibility of the MT output.

3.2 Checking MWE Translation Quality

The question arise if the automatic translation of MWEs requires any correction, in the first place. As Babych observes, "SMT output is often surprisingly good with respect to short distance collocations" (Babych et al., 2012, 103). Good translations for idiomatic expressions can still be achieved in SMT as a by-product of learning from parallel corpora. This can be seen, for instance, in Example (4)). The MT output required no correction at all from the post-editor.

(4) a. Faites-nous part de vos expériences
b. Please email us your experiences

Example (5), on the contrary, shows a bad translation. The collocation *rencontrer erreur* is translated literally by the system.

(5) a. *Source:* Je viens de *rencontrer* une *erreur* à l 'instant en faisant un Live Update manuel
b. *MT:* I have just *met* an *error* just now by a Live Update manual
c. *Post-editor:* I have just *had* an *error* just now doing a manual Live Update

To report the number of MWEs that are well translated by the system, we relied on the post-editor's version as a gold standard. Whenever the editor changed the MWE translation as proposed by the system, we considered it was wrong, except for the cases where the changes were minor, like fixing number or agreement.

According to this method, the percentage of well-translated MWEs is 63.2% (141/223). Therefore, less than half of MWEs required a different translation. This result contradicts our expectations induced by theoretical claims, which would predict a higher rate of failure. It might also explain the limited impact of MWE integration observed in the literature: if MWEs account for about 5% of the data and more than half are well translated anyway, the small increase in BLEU seems justified.

Previous research (Bod, 2007; Wehrli et al., 2009; Babych et al., 2012) has suggested that SMT is more problematic for the more flexible expressions. This problem is exacerbated in our domain, as shown in Example (6). The SMT system fails to correctly translate the MWE *mise à jour* because its form deviates from the expected form and takes an unconventional plural form:

(6) a. *Source: mise à jours* live update
b. *MT: Upgrade days* live update
c. *Post-editor: Update* to live Update .

Due to time constraints, for the present experiment we did not relate yet the quality of MWE translation to the flexibility of the expressions, in order to find wether there is an effect. This analysis is left for future work. We tested, however, the statistical significance of the difference between the total number of MWE in the 152 segments containg MWEs, on the one hand, and the number of correctly translated MWEs, on the other hand. This difference is extremely significant ($t(151) = 9.93, p < 0.001$). This means that the problem of MWEs in translation is real. A significant number of MWEs are badly translated. If we focus on MWEs, we only deal with about 10% of the data (see Section 3.1), but arguably we deal with the most critical portion of the data compared to other corrections which might not be as critical. Fixing a determiner, number or agreement might not have the same impact on comprehensibility as fixing a collocate (see Example (5)). Moreover, MWE translation errors seem to make a large share of all errors, because MWEs are common and they are often badly translated. This hypothesis is tested in the experiment described next.

	BLEU	WER	TER	Levenshtein
Total effort (MT)	0.511	0.316	0.291	24.0
Effort excluding MWE correction (Oracle)	0.603	0.249	0.225	19.8
Effort spent on MWEs (difference)	0.092	-0.066	-0.066	4.2
Effort spent on MWEs (%)	18.0%	-21.0%	-22.6%	17.6%
t(151)	-6.83	12.25	6.16	8.46

Table 2: Post-editing effort spent fixing MWE translation errors.

3.3 Quantifying MWE Correction Effort

How much of the total post-editing effort is actually spent on fixing MWE translation errors? To answer this question, we quantified the post-editing effort in terms of standard metrics used in the field (BLEU, TER, WER, Levenshtein).[4] We compared the total post-editing effort against the post-editing effort excluding MWE correction. The difference represents the effort devoted to fixing MWE translation errors.

The total post-editing effort is, obviously, computed for the MT output as such. The effort excluding MWE correction is computed on a modified version, on which the correct, 'oracle' MWE translation is extracted from the gold standard, which is the post-editor's version. To illustrate this, we provide an example below (Example (7)). The MWE correction *tried all ways* is inserted from the post-editor's version, while the rest is left unchanged (notice the post-editor further changed *do not* into *can't* at the end of the sentence).

(7) a. *Source:* J'ai *retourné* le programme *dans tout les sens* pour trouver l'option qui permet de changer le mot de passe mais je ne la trouve pas.

b. *MT:* I have *returned* the program *in any sense* to find the option that lets you change the password but I do not find it.

c. *Post-editor:* I have *tried all ways* to find the option that lets you change the password but I can't find it.

d. *Oracle:* I have *tried all ways* to find the option that lets you change the password but I do not find it.

The results are shown in Table 2. MWEs account for about a fifth of the total post-editing effort, according to the metrics used. Admittedly, this is less than expected considering theoretical arguments. Like the bad translation hypothesis, the hypothesis that most of the post-editing effort is focused on MWEs is invalidated by our study. However, this result is based on the selection of metrics used to quantify the post-editing effort. We believe that there are more accurate metrics of measuring effort, like time. Had we had time logs for our data, we could have come up with a different conclusion. Indeed, the time needed for providing a correct translation for a MWE is arguably much longer than the time required to delete a determiner of to fix agreement issues. Further investigation is therefore needed in order to reliably invalidate the hypothesis in question.

As for the statistical significance of results, the difference between the total effort and the effort excluding MWEs correction is extremely significant ($p < 0.001$), as can be seen in the last row of Table 2. This means that the MWE correction effort is significant. Again, the interpretation of this finding is that MWEs constitute a real problem for machine translation.

It is important to note that if MWEs are handled perfectly, the expected increase in translation quality can be as high as 9.2 BLEU points, while current integration methods achieve about 1 BLEU point, as seen in Section 1.

4 Conclusion

Summing up, while the literature put emphasis on the prevalence on the prevalence of MWEs and their importance for translation, little was known about the empirical validity of theoretical claims, and even less so about their validity in the specific domain of user-generated content. This domain is little investigated by MWE research, but is of major interest for natural language processing in general and for machine translation in particular.

[4] BLEU measures the distance from a reference translation at the word level, using n-grams. TER measures the same distance in terms of operations at the word level (substitution, insertion, deletion, shift). WER is similar to TER, with no shift. The Levenshtein distance on which TER and WER are based works similarly, but at the character level. Other post-editing effort measures are the time and keystrokes. Time and keystoke logs are unfortunately not available for our data.

The aim of the present study was to test the validity of theoretical claims for this domain, in order to find out, in particular, how frequent MWEs and MWE translation errors really are, and how much of the total post-editing effort is spent on correcting MWE translation errors. We conducted a study based on large-scale post-editing dataset, which allowed us to validate the MWE prevalence assumption and to find out that MWEs account for more than 10% of words in our dataset. We also checked the bad translation assumption and found that the majority of MWEs are actually correctly translated. This is different from what the literature suggests, but we found that the number of badly-translated MWEs is, however, significant. As for the integration of MWE knowledge into MT systems, we computed an upperbound for the increase in translation quality we could expect by better handling MWEs: if we handle them perfectly, we could gain as much as 9 BLEU points. These results suggest that there is still room for improvement in this area.

This study could be extended to more language pairs and new datasets, by exploiting multiple annotations, and quantifying the MWE translation correction effort in terms of time, in addition to automatic metrics.

References

Bogdan Babych, Kurt Eberle, Johanna Geiß, Mireia Ginestí-Rosell, Anthony Hartley, Reinhard Rapp, Serge Sharoff, and Martin Thomas. 2012. Design of a hybrid high quality machine translation system. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 101–112, Avignon, France, April.

Ming-Hong Bai, Jia-Ming You, Keh-Jiann Chen, and Jason S. Chang. 2009. Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 478–486, Singapore.

Rens Bod. 2007. Unsupervised syntax-based machine translation: the contribution of discontiguous phrases. In *Proceedings of MT Summit XI*, pages 51–56, Copenhagen, Denmark, September.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California, June. Association for Computational Linguistics.

2013. ACCEPT deliverable D 4.1: Baseline MT systems. `http://www.accept.unige.ch/Products/D_4_1_Baseline_MT_systems.pdf`.

Peter Howarth and Hilary Nesi. 1996. The teaching of collocations in EAP. Technical report, University of Leeds, June.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Valia Kordoni and Iliana Simova. 2014. Multiword expressions in machine translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Diana Lea and Moira Runcie, editors. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford University Press, Oxford.

Zhanyi Liu, Haifeng Wang, Hua Wu, and Sheng Li. 2010. Improving statistical machine translation with monolingual collocation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 825–833, Uppsala, Sweden, July.

Zhanyi Liu, Haifeng Wang, Hua Wu, Ting Liu, and Sheng Li. 2011. Reordering with source language collocations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1036–1044, Portland, Oregon, USA, June.

Kathleen R. McKeown and Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*, pages 507–523. Marcel Dekker, New York, USA.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony P. Cowie, editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Claredon Press, Oxford.

Brigitte Orliac and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX*, pages 292–298, New Orleans, Lousiana, USA.

Darren Pearce. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 41–46, Pittsburgh, USA.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China, August.

Eric Wehrli, Violeta Seretan, Luka Nerima, and Lorenza Russo. 2009. Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, pages 128–135, Barcelona, Spain.