TED-MWE: a bilingual parallel corpus with MWE annotation

Towards a methodology for annotating MWEs in parallel multilingual corpora

Johanna Monti[1]*, Federico Sangati[2], Mihael Arcan[3]
[1]Sassari University, Sassari, Italy
[2]Fondazione Bruno Kessler, Trento, Italy
[3]National University of Ireland, Galway, Ireland
jmonti@uniss.it,sangati@fbk.eu,mihael.arcan@insight-centre.org

Abstract

English. The translation of Multiword expressions (MWE) by Machine Translation (MT) represents a big challenge, and although MT has considerably improved in recent years, MWE mistranslations still occur very frequently. There is the need to develop large data sets, mainly parallel corpora, annotated with MWEs, since they are useful both for SMT training purposes and MWE translation quality evaluation. This paper describes a methodology to annotate a parallel spoken corpus with MWEs. The dataset used for this experiment is an English-Italian corpus extracted from the TED spoken corpus and complemented by an SMT output.

Italiano. *La traduzione delle polirematiche da parte dei sistemi di Traduzione Automatica (TA) rappresenta un sfida irrisolta e benché i sistemi abbiano compiuto notevoli progressi, traduzioni errate di polirematiche occorrono ancora molto di frequente. È necessario sviluppare ampie collezioni di dati principalmente corpora paralleli annotati con polirematiche che siano utili sia per l'addestramento della TA di tipo statistico sia per la valutazione della qualità della traduzione delle polirematiche. Questo contributo descrive una metodologia per annotare un corpus parallelo del parlato con le polirematiche e il corpus stesso. La collezione di dati usata per questo esperimento è un corpus inglese-italiano estratto dal TED, corpus del parlato, integrato dalla traduzione di un sistema statistico di TA.*

1 Introduction

Multiword expressions (MWEs) represent one of the major challenges for all Natural Language Processing (NLP) applications and in particular for Machine Translation (MT) (Sag et al., 2002). The notion of MWE includes a wide and frequent set of different lexical phenomena with their specific properties, such as idioms, compound words, domain specific terms, collocations, Named Entities or acronyms. Their morpho-syntactic, semantic and pragmatic idiomaticity (Baldwin and Kim, 2010) together with translational asymmetries (Monti and Todirascu, 2015), i.e. the differences between an MWE in the source language and its translation, prevent technologies from using systematic criteria for properly handling MWEs. For this reason their automatic identification, extraction and translation are very difficult tasks.

Recent PARSEME surveys[1] have highlighted that there is lack of MWE-annotated resources, and in particular parallel corpora. Moreover, the few available ones are usually limited to the study of specific MWE types and specific language pairs. The focus of our research work is therefore to provide a methodology for annotating a parallel corpus with all MWEs (with no restrictions to a specific type) which can be used both for training and testing SMT systems. We have refined this methodology while developing the English-Italian MWE-TED corpus, which contains 1.5K sentences and 31K EN tokens.It is a subset of the TED spoken corpus annotated with all the MWEs detected during the annotation process. This contribution presents the corpus[2] together with the annotation guidelines in section 3, the annotation process in section 4 and the MWE annotation statistics in section 5.

[*] Johanna Monti is author of sections 2 and 3.2, Federico Sangati is author of sections 4 and 5, Mihael Arcan is author of sections 3.1 and 4.1. Introduction and conclusions are in common.

[1] Translating Multiword Expressions - PARSEME WG3 State of the Art Report - forthcoming

[2] http://tiny.cc/TED_MWE

2 Related work

As mentioned in the previous section, the research work in this field is mainly focused on the annotation of specific MWE types, such as (i) the SzegedParalell English-Hungarian parallel corpus (Vincze, 2012) which contains 1370 occurrences of light verb constructions (LVCs), (ii) 4FX, a quadrilingual parallel corpus annotated manually for LVCs (Rácz et al., 2014) containing 673 LVCs in English, 806 in German, 938 in Spanish and 1059 in Hungarian.

Unlike the above methodologies, our aim is to provide a more general approach to MWE annotation in a parallel and multilingual corpus. In this respect, Schneider et al. (2014) present an interesting *comprehensive annotation approach*, in which all different types of MWEs are annotated in a 55K-word corpus of English web text.

Annotating MWEs in parallel texts involves several problems due to the translational asymmetries between languages and presence of discontinuity, but it is considered very important to compensate for the lack of training and benchmark resources for MT.

There are few corpora specifically built to evaluate MT translation quality with reference to MWE translation, such as (i) Ramisch et al. (2013) where an English-French corpus annotated with Phrasal Verbs (PVs) is used to assess the quality of PV translation by a phrase-based system (PBS) and a hierarchical system (HS) or (ii) Schottmüller and Nivre (2014), who describe a German-English corpus containing Verb-particle constructions (VPCs), used to compare the results obtained from Google Translate and Bing Translate, and finally Barreiro et al. (2013), who use parallel corpora (English to Italian, French, Portuguese, German and Spanish) containing 100 English Support Verb Constructions (SVC) and their translations in the target languages done by OpenLogos and the Google Translate.

3 TED-MWE

3.1 The TED Corpus

We have used the WIT^3 web inventory (Cettolo et al., 2012) which offers access to a collection of transcribed and translated talks. The core of WIT^3 is the TED Talks corpus, that basically redistributes the original content published by the TED Conference website. The WIT^3 corpus repurposes the original TED content in a way which is more convenient for MT researchers. For our experiments we used the WIT^3 data released for the IWSLT 2014 Evaluation Campaign, which contains the training data of 190K parallel sentences, needed to build an SMT system. We base our annotations and analysis on the test set, which we will refer to as the MWE-TED corpus.

3.2 MWE Annotation Guidelines

The judgement of whether an expression should qualify as an MWE relies on the annotation guidelines, which are based on the PARSEME MWE template and the testing of MWE properties.

The PARSEME MWE Template provides information and examples for all different MWE syntactic structures (nominal verbal, adjectival, prepositional, clausal MWEs), the fixedness/flexibility of MWE parts, the different levels of idiomaticity (lexical, syntactic, semantic, pragmatic, statistical idiomaticity) and finally the rhetoric relations within an MWE. In addition to the template, annotators were provided with a set of tests (Monti, 2012) to be used to assess if a certain group of words can be considered as a MWE:

Non-substitutability : one element of the MWE cannot be replaced without a change of meaning or without obtaining a non-sense (*in deep water* → *in hot water*; *gas chamber* → **gas room*);

Non-expandability : insertion of additional elements is not possible (*get a head start* → **get a quick head start*);

Non-reducibility : the elements in the MWE cannot be reduced and pronominalisation of one of the constituents is also not possible (*take advantage* → **what did you take? advantage*; **Did you take it?*;

Non-literal translatability : the meaning cannot be translated literally. The difficulty of a literal translation across cultural and linguistic boundaries is mainly a property of MWEs with limited or no variation of distribution, such as idioms (e.g., *it's raining cats and dogs* → it. **sta piovendo cani e gatti*), but also of many collocations (e.g., *heavy rain* → it. **pioggia pesante*), fixed expressions (e.g., *by and large* → it. **da e largo*), proverbs (e.g., *there's no such thing as a free lunch* → it. **non esiste una cosa come un pranzo gratuito*), phrasal verbs (e.g., *bring somebody down* → it. **Portare qualcuno giù*);

Invariability : Invariability can affect both the morphological and the syntactic level. Inflectional variations of the constituents of the MWEs are not always possible. Invariability affects both the head elements and its modifiers (*fish out of water* → **fishes out of water*; *dead on arrival* → **dead on arrivals*; *in high places* → **in high place*); syntactical variations inside an MWU may also not be acceptable (*credit card* → **card of credit*);

Non-displaceability : displacement and a different order of constituents are not possible (*wild card* → **is wild this card?*) - (*back and forth* → **forth and back*);

Institutionalisation of use : certain word units, even those that are semantically and distributionally ”free”, are used in a conventional manner. The Italian expression *in tempo reale* (a loan translation of the English expression *in real time*) is an example of this feature since its antonym **in tempo irreale* (**in unreal time*) seems to be unmotivated and not used at all.

In order to consider a certain word unit as an MWE it is sufficient that it shows at least one of the above-mentioned properties. Nevertheless, during the annotation process, the property which turned out to characterise the majority of MWEs is the non-literal translatability.

4 Annotation Process

The annotation was organised in three distinct phases: individual annotation, inter-annotation check, validation.

Individual annotation. During the first phase, thirteen annotators with linguistic background in Italian and English were asked to annotate the 1,529 sentences in the MWE-TED corpus. The sentences were organised in a spreadsheet (see figure 1) containing the following information: (i) the English source text, (ii) the Italian *manual* translations (from the parallel corpus) and finally (iii) the Italian *SMT* output (see section 4.1). The annotators were asked to identify all the MWEs in the source text together with their translations in approximately 300 random sentences each and to evaluate the automatic translation correctness[3]. If the *manual* or the *SMT* generated translations were wrong, the annotators were asked to specify the correct translations.

The annotation took into account all MWE types detected in the source text with no restrictions to a particular type of MWE and in particular, both contiguous and discontinuous MWE types were recorded in the dataset. The MWEs identified during the annotation process were recorded as sequences of tokens with no further information about their internal syntactic structure or semantic features.

Inter-annotation check. In the second phase, each annotator was confronted with the anonymized annotations by the other annotators on his/her annotation subset, in order to decide about his/her choices, i.e. to confirm or change the annotations for each source text/manual/SMT set (see table 1).

Sentence: 369
Source: people sort of think i went away between “ titanic ” and “ avatar ” and was buffing my nails someplace , sitting at the beach .
Your MWE(s) [sort of, buffing my nails, someplace]
Ann.10 MWE(s) [sort of, buffing my nails]

Sentence: 432
Source: now that ’s back from high school algebra , but let ’s take a look .
Your MWE(s) [back from]
Ann.6 MWE(s) [take a look]

Sentence: 539
Source: that ’s a key element of making that report card .
Your MWE(s) [report card]
Ann.12 MWE(s) [key element, report card]

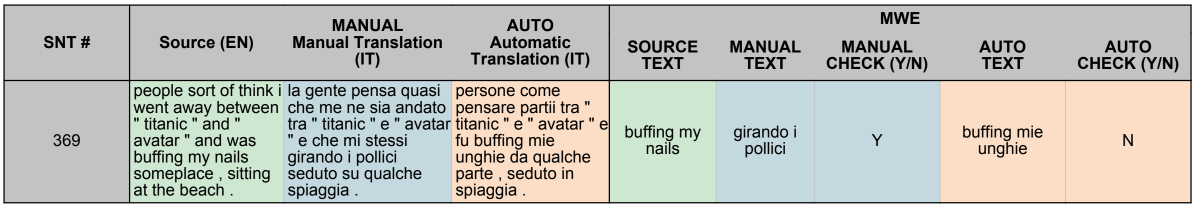
Table 1: Annotation phase 2: inter-annotation check.

Validation. Finally, in the last phase, we have randomly selected about half of the annotated sentences (801) and asked the annotators to integrate and resolve the possible annotation conflicts (see figure 2).

4.1 Statistical Machine Translation

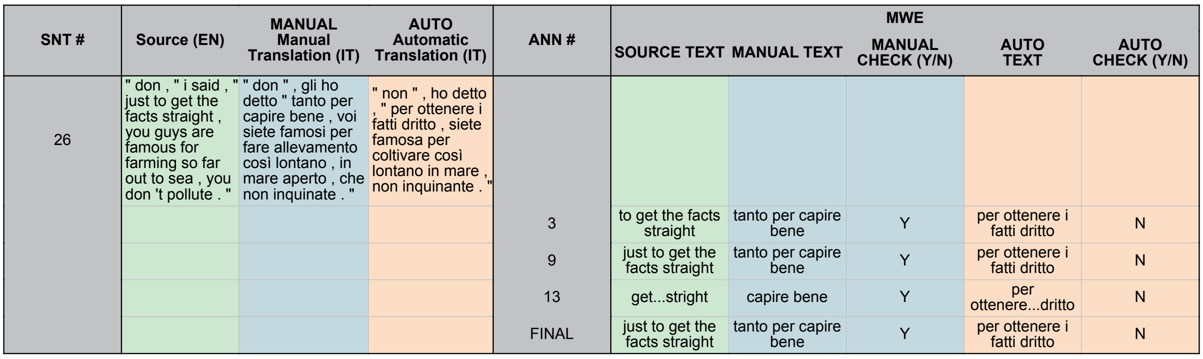
In order to gather automatic translations of the source text, we used the Moses toolkit (Koehn et al., 2007), where the word alignments were built with GIZA++ (Och and Ney, 2003). The IRSTLM toolkit (Federico et al., 2008) was used to build the 5-gram language model. The parameters within the SMT system are optimized on the development data set using MERT (Bertoldi et al., 2009). The system performed in line with the state-of-the-art results on the test set.

[3] The annotation work was organised in such a way that each sentence was annotated by at least two annotators



SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	MWE				
				SOURCE TEXT	MANUAL TEXT	MANUAL CHECK (Y/N)	AUTO TEXT	AUTO CHECK (Y/N)
369	people sort of think i went away between " titanic " and " avatar " and was buffing my nails someplace , sitting at the beach .	la gente pensa quasi che me ne sia andato tra " titanic " e " avatar " e che mi stessi girando i pollici seduto su qualche spiaggia .	persone come pensare partii tra " titanic " e " avatar " e fu buffing mie unghie da qualche parte , seduto in spiaggia .	buffing my nails	girando i pollici	Y	buffing mie unghie	N

Figure 1: Annotation phase 1: individual annotation.



SNT #	Source (EN)	MANUAL Manual Translation (IT)	AUTO Automatic Translation (IT)	ANN #	MWE				
					SOURCE TEXT	MANUAL TEXT	MANUAL CHECK (Y/N)	AUTO TEXT	AUTO CHECK (Y/N)
26	" don , " i said , " just to get the facts straight , you guys are famous for farming so far out to sea , you don 't pollute . "	" don " , gli ho detto " tanto per capire bene , voi siete famosi per fare allevamento così lontano , in mare aperto , che non inquinate . "	" non " , ho detto , " per ottenere i fatti dritto , siete famosa per coltivare così lontano in mare , non inquinante . "						
				3	to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				9	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N
				13	get...stright	capire bene	Y	per ottenere...dritto	N
				FINAL	just to get the facts straight	tanto per capire bene	Y	per ottenere i fatti dritto	N

Figure 2: Annotation phase 3: validation

English	Italian
pointed at	indicò
no longer	non ... più
don 't get me wrong	non fraintendetemi
got bitten by	sono stato affetto dal
a lot of	un sacco di
in the dead of winter	nella tristezza dell' inverno

Table 2: Sample of annotated MWE EN-IT pairs.

5 MWE Annotation Statistics

After the first two phases of the annotation process, out of 1,529 annotated sentences, 541 (35.9%) showed a good inter-annotation agreement, i.e. at least two annotators completely agreed on the annotations. In total we have collected 2,484 English MWEs types out of which 2,391 (96%) are contiguous and 93 (4%) are discontinuous. At least two annotators agreed for the 27% (671) of the MWEs and in 45% of them (1,115) at least two annotators showed an overlapping (at least one word in common).

This general low agreement scores confirm the difficulty of the annotation task. In order to resolve the numerous annotation conflicts, we ran a third annotation phase in which 801 of the previous sentences were validated. This resulted in a total of 799 English MWE types (931 tokens), of which 729 (91%) are contiguous and the 9% (70) are discontinuous. Most MWEs have length 2 (515) and 3 (261), but there are MWEs up to length 8. In 52% of the cases (471) the annotators have evaluated the automatic translation to be incorrect. Table 2 reports a small sample of annotated English MWEs together with their Italian translations.

6 Conclusions

We have described the TED-MWE corpus, an English-Italian parallel spoken corpus annotated with MWEs, together with the methodology and the guidelines adopted during the annotation process. Ongoing and future work includes refinement of the annotation tools and guidelines, the extension of the methodology to further languages in order to develop a multilingual MWE-TED corpus. The main aim is to provide useful data both for SMT training purposes and MT quality evaluation.

Acknowledgments

We greatly acknowledge the PARSEME IC1207 COST Action for supporting this work. We are particularly grateful to Manuela Cherchi, Erika Ibba, Anna De Santis, Giuseppe Casu, Jessica Ladu, Ilaria Del Rio, Elisa Virdis, Gino Castangia for their annotation work.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 1, pages 267–292. CRC Press, Boca Raton, USA, second edition edition.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When multiwords go bad in machine translation. *MT Summit workshop Proceedings on Multi-word Units in Machine Translation and Transla tion Technology*, page 10.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved minimum error rate training in moses. *Prague Bull. Math. Linguistics*, 91:7–16.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268. Trento, Italy.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, pages 1618–1621.

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. Association for Computational Linguistics, Prague, Czech Republic.

Johanna Monti. 2012. *Multi-word unit processing in Machine Translation - Developing and using language resources for Multi-word unit processing in Machine Translation*. Ph.D. thesis, University of Salerno.

Johanna Monti and Amalia Todirascu. 2015. Multiword Units Translation Evaluation: another pain in the neck? In *Proceedings of Multi-word Units in Machine Translation and Translation Technology (MUMTTT15)*. Malaga.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.

Anita Rácz, István Nagy T., and Veronika Vincze. 2014. 4fx: Light verb constructions in a multilingual parallel corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

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

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461. European Language Resources Association (ELRA), Reykjavik, Iceland.

Nina Schottmüller and Joakim Nivre. 2014. Issues in translating verb-particle constructions from german to english. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 124–131. Association for Computational Linguistics, Gothenburg, Sweden.

Veronika Vincze. 2012. Light verb constructions in the szegedparalellfx english–hungarian parallel corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.