Discovering Light Verb Constructions and their Translations from Parallel Corpora without Word Alignment

Natalie Vargas [1], Carlos Ramisch [2] and Helena de M. Caseli [1]

[1] Federal University of São Carlos, São Carlos, Brazil
{helenacaseli, natalie.vargas}@dc.ufscar.br

[2] Aix Marseille Univ, CNRS, LIF, Marseille, France
carlos.ramisch@lif.univ-mrs.fr

Abstract

We propose a method for joint unsupervised discovery of multiword expressions (MWEs) and their translations from parallel corpora. First, we apply independent monolingual MWE extraction in source and target languages simultaneously. Then, we calculate translation probability, association score and distributional similarity of co-occurring pairs. Finally, we rank all translations of a given MWE using a linear combination of these features. Preliminary experiments on light verb constructions show promising results.

1 Introduction

The automatic discovery of multiword expressions (MWEs) has been a topic of interest in the computational linguistics community for a while (Choueka, 1988; Church and Hanks, 1990). In the last 20 years, *multilingual* discovery of MWEs has gained some popularity thanks to the widespread use of statistical machine translation (MT), automatic word alignment tools and freely available parallel corpora (Zarrieß and Kuhn, 2009; Attia et al., 2010; Caseli et al., 2010). MWEs tend to be non compositional or show some kind of lexico-syntactic inflexibility, which is often reflected in translation asymmetries (Manning and Schütze, 1999). Therefore, parallel corpora are rich resources to mine for MWEs. Techniques adapted from machine translation can help to exploit translation information for the specific needs of MWE discovery.

Parallel corpora can be useful for MWE discovery in many ways. First, a second (target) language can be used to model features, which in turn help in the discovery of new MWEs in a single (source) language (Salehi and Cook, 2013; Caseli et al., 2010; Tsvetkov and Wintner, 2014). Second, one can also use parallel data to discover the translations of known multiword lexical units (Morin and Daille, 2010). Finally, it is possible to perform both simultaneously, generating a bilingual lexicon of MWEs and their potential translations from the parallel corpus, as proposed in this paper.

The goal of our paper is to propose a new method for unsupervised joint discovery of MWEs and their translations. It consists in discovering potential MWEs on source and target texts independently, and then trying to match them without using automatic word alignment. It is important to emphasize that we are not against the use of word alignment for this task, but we are interested in seeing how the automatic discovery of MWEs can be performed without relying on this information. Moreover, our experiments focus on light verb constructions such as *to make a presentation* and *to take a walk*, which generally contain non-adjacent tokens and thus would probably not be captured by standard word alignment methods. We study several features to rank automatically extracted candidates that could be translations of each other. We show preliminary results that indicate this approach is promising and point towards future improvements.

2 Related Work

Multilingual resources in general can be used for MWE discovery. Attia et al. (2010), for instance, do not rely on parallel texts but on short Wikipedia page titles, cross-linked across multiple languages. They consider that, if a page whose title contains a cross-lingual link to a page whose title is a single word (in any available language), then the original page title is probably a MWE. Similarly, translation links in Wiktionary can be exploited, among

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 91–96,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

other features, for predicting the compositionality of MWEs (Salehi et al., 2014a).

Another possibility to model non-translatability without recurring to parallel corpora consists in building up artificial word-for-word MWE translations using bilingual single-word dictionaries. Afterwards, the existence of these automatically generated potential translations can be assessed in large monolingual corpora (Morin and Daille, 2010). This can be used as a feature, among other sources of information, in supervised or semi-supervised monolingual MWE discovery (Tsvetkov and Wintner, 2011; Rondon et al., 2015). Bilingual dictionaries can also be used to predict the compositionality of MWEs by estimating the string similarity (Salehi and Cook, 2013) or distributional similarity (Salehi et al., 2014b) between translations of an MWE and of the single words it contains.

Melamed (1997) describes one of the earliest attempts to extract MWEs from parallel corpora. The method is based on lexical alignment and mutual information. Statistical lexical alignment can provide straightforward MWE candidates, which can be further filtered using POS patterns and association scores. If two or more words in a source language are aligned to the same word on the target side, the source is likely an MWE (Caseli et al., 2010). Conversely, one can assume that some types of MWEs such as verb-noun combinations tend to be translated as MWEs with the same syntactic structure, using aligned dependency-parsed corpora for discovery (Zarrieß and Kuhn, 2009). Instead of focusing on 1-to-many alignments, Tsvetkov and Wintner (2010) propose a method which incrementally removes from parallel sentences word pairs that are surely not MWEs. Therefore, they use bilingual dictionaries and alignment reliability scores. The remaining units are considered candidate MWEs.

Bilingual lexicons containing MWEs are important resources for MT systems. It has been shown that the presence of MWEs can harm the quality of both statistical (Ramisch et al., 2013) and rule-based (Barreiro et al., 2014) MT systems. Simple techniques for taking MWEs into account such as binary features (Carpuat and Diab, 2010) and special token markers (Cap et al., 2015) can help improving translation quality. However, this may not suffice if the expressions are not correctly identified with the help of bilingual MWE lexicons.

3 Bilingual MWE Lexicon Creation

Most existing methods exploit parallel corpora to discover MWEs in a single language. They use translation information, among other sources, to confirm the idiosyncratic behaviour of the MWE in the source language, but do not output possible translations as a result of the discovery algorithm. In this section, we propose a method to create probabilistic bilingual MWE dictionaries using minimal supervision.

First, we extract MWE candidates from pre-processed (POS-tagged and lemmatized) source and target texts separately. In our experiments, the texts were pre-processed by TreeTagger (Schmid, 1994). We explicitly configured it not to segment sentences, since we need to preserve the alignment between source and target sentences in our input parallel corpus.

To allow the extraction of these monolingual MWE candidates, it is necessary to manually define POS patterns in both languages. This step requires some knowledge about the languages and about the syntactic patterns of the MWEs that we want to extract. These patterns were defined using the mwetoolkit corpus query language and candidate extraction tools (Ramisch, 2015).[1] In this first moment, we focused on MWEs translated into MWEs, but we believe that the technique could be adapted to MWEs translated into single words. For instance, one could extract verbal MWEs from the source corpus and try to match them with single-word verbs in the target language. In theory, any monolingual MWE discovery approach could be used to obtain candidates on each side of the parallel corpus independently.

The process described above outputs two sets of candidates. The first set $S = \{s_1, s_2, \ldots, s_{|S|}\}$ contains MWE candidates s_i extracted from the source corpus. The second set $T = \{t_1, t_2, \ldots, t_{|T|}\}$ contains MWE candidates t_j extracted from the target corpus. Then, we try to map source MWEs s_i to their target correspondences t_j. To do so, we calculate the **conditional probability** of each potential translation (t_j) in T given a source (s_i):

$$P(t_j|s_i) = \frac{c(s_i, t_j)}{c(s_i)}$$

Here, $c(s_i, t_j)$ is the number of times a source candidate s_i was found in a sentence whose transla-

[1] http://mwetoolkit.sf.net/

tion contained t_j and $c(s_i)$ is simply the number of occurrences of the candidate in the source corpus. Since candidates s_i and t_j can be discontinuous, their numbers of occurrences are not necessarily n-gram counts, but must be obtained during monolingual candidate discovery as output by the mwetoolkit.

Another measure that we use to rank translations is the **t-score**. This association score estimates to what extent the co-occurrence of a group of words is outstanding compared to random chance co-occurrence. For each target candidate $t_j = w_1^{t_j} w_2^{t_j} \ldots w_n^{t_j}$, formed by n words $w_k^{t_j}$, we compute the expected number of occurrences by multiplying all individual word probabilities $\frac{c(w_k^{t_j})}{N}$ and then scaling this joint probability by the total number of tokens in the target corpus N:

$$E(t_j) = \frac{c(w_1^{t_j}) \times c(w_2^{t_j}) \times \ldots \times c(w_n^{t_j})}{N^{n-1}}$$

The t-score, also obtained using the mwetoolkit, is the difference between observed and expected counts normalized by an estimate of the standard deviation of the distribution:

$$tscore(t_j) = \frac{c(t_j) - E(t_j)}{\sqrt{c(t_j)}}$$

Finally, we calculate the multilingual distributional **similarity** between pairs s_i and t_j. This score is based on a pre-trained vector space model which uses sentence alignment information to ensure that words that are translations of each other end up being close in the resulting semantic space. Since each unit s_i and t_j is composed of m and n words, respectively, we use the average cosine similarity between all possible $m \times n$ source-target pairs present in the semantic space:[2]

$$Sim(s_i, t_j) = \frac{1}{m \times n} \sum_{\substack{k = 1..m \\ l = 1..n}} cos(w_k^{s_i}, w_l^{t_j})$$

The bilingual semantic space is obtained using MultiVec (Bérard et al., 2016).[3] Distributional similarity between source and target candidate words is obtained using the *bag of words* mode.

The three scores are normalized so that their values fall between 0 and 1. The final score F is simply a log-linear combination of these scores:

$$F(t_j|s_i) = \sum_{f \in \{P, tscore, Sim\}} - \log norm(f(t_j, s_i))$$

The lower its value, the more likely a given pair of source and target MWEs is.

4 Experimental Setup

For this work, the pre-processed texts (POS-tagged source and target texts) were obtained from the FAPESP parallel corpus containing 166,719 aligned sentences of Brazilian Portuguese texts translated into English (Aziz and Specia, 2011). The source corpus contains 4,191,942 tokens and the target corpus contains 4,499,064 tokens.[4]

Our experiments employ manually defined patterns for the monolingual step. These patterns target light-verb constructions in Portuguese and some possible translations into English:

GET+ADJ The first pattern consists of the Portuguese verb *ficar* (*to become*) immediately followed by an adjective. This frequent construction often indicates a change of state (inchoative). On the target language (English), we build a similar pattern consisting of verbs *to be/become/get* + an adjective, which we assume as being frequent translations for the source construction.

MAKE+N This pattern is formed by the verb *realizar* (*to make*) followed by a noun. Between the verb and the noun there can be any number of adjectives, adverbs or determinants, which are ignored in the extracted candidate. For the translation, we build an equivalent pattern with verbs *to make/carry* due to the high occurrence of *carry out* in the target corpus.

TAKE+N This pattern is formed by verbs *fazer/tomar/dar* (*to make/take/give*) followed by a noun. We allow intervening elements as for MAKE+N. In English, we use verbs *to make/do/take*. Notice that verb *to give* was considered as an unlikely translation and disregarded.

5 Preliminary Results

As mentioned in Section 3, we used the mwetoolkit to apply the patterns and calculate t-scores

[2] The normalization factor may be less than $m \times n$ when some pairs $w_k^{s_i}, w_l^{t_j}$ do not occur in the semantic space.

[3] `https://github.com/eske/multivec`

[4] `http://www.nilc.icmc.usp.br/nilc/tools/Fapesp\%20Corpora`

	MWE source	MWE target	# T	*ts* T	Sim	F
1	**ficar doente**	**get sick**	2	0.51	0.53	1.44
2	**ficar doente**	**become ill**	2	0.50	0.46	1.51
3	ficar doente	be normal	1	0.52	0.41	1.84
4	**ficar doente**	**become sick**	1	0.49	0.41	1.86
5	ficar doente	be tolerant	1	0.50	0.33	1.95
1	**ficar pronto**	**be ready**	46	0.72	0.67	0.41
2	**ficar pronto**	**become ready**	5	0.50	0.60	1.58
3	**ficar pronto**	**get ready**	1	0.58	0.69	2.15
4	ficar pronto	be capable	2	0.74	0.25	2.18
5	ficar pronto	be necessary	1	0.87	0.40	2.22
6	ficar pronto	be fundamental	1	0.74	0.26	2.47

Table 1: Pattern GET+ADJ: *ficar doente/pronto* (*get sick/ready*). Correct pairs are in bold.

and MultiVec for bilingual similarity. Unfortunately, quantitative evaluation was not yet performed. Nonetheless, in this section, we present some examples of discovered MWEs along with their translations. We point out positive and negative results in this small sample that give us an idea of our approach's potential.

Table 1 shows ranked examples extracted from the source and target corpus for the first pattern. The entries are ranked by final score, more likely translations appear on the top of the table and the correct ones are in bold. According to these examples, the MWE pairs with lowest scores are correctly aligned to a valid translation. In addition to the final score (F), target t-score (*ts* T) and similarity (Sim), the table also shows how many times the source MWE co-occurred with the target MWE (# T). This information allows us to calculate the conditional probability.

It is important to point out that our approach does not work for all cases, as some spurious pairs also occur. For example, in the first half of table 1, *become sick* is indeed a possible translation for *ficar doente* but it appears in a worst position compared to *be normal*, which is not a possible translation. Beyond the conditional probability, distributional similarity and t-score seem to help in some cases. For instance, *get ready* appears only once as a translation of *ficar pronto*, but still it gets a better score than *be capable*, a wrong translation with higher conditional probability. In general, we have observed that the pattern GET+ADJ is quite "easy" to translate as these constructions show a high degree of regularity.

	MWE source	MWE target	# T	*ts* T	Sim	F
1	**realizar teste**	**carry test**	20	0.50	0.73	0.71
2	**realizar teste**	**carry trial**	3	0.29	0.63	1.85
3	realizar teste	carry field	4	0.26	0.47	1.89
4	realizar teste	make assessment	4	0.22	0.28	2.18
5	realizar teste	make use	1	1.00	0.24	2.19
6	**realizar teste**	**make test**	1	0.23	0.62	2.43
7	realizar teste	make comparison	1	0.38	0.30	2.51
8	**realizar teste**	**carry test**	1	0.17	0.65	2.54
9	realizar teste	carry safety	1	0.17	0.53	2.64
10	realizar teste	make prototype	1	0.23	0.37	2.64
11	realizar teste	make search	1	0.15	0.24	3.02
1	realizar substituição	carry identification	1	0.19	0.45	1.67

Table 2: Pattern MAKE+N: *realizar teste/substituição* (*make test/replacement*). Correct pairs are in bold.

Table 2 shows the results of the extraction for MAKE+N. The results for *realizar teste* show that the best ranked MWEs are the corrected translations. The last row of this table shows a drawback of our approach: that it is not possible to obtain reliable probability scores when the pattern just appears once.

The results in table 3 show the extraction for the last pattern, TAKE+N. Despite the first half of this table presenting good results for *do comparison* and *make comparison*, the second half shows that some patterns do not work for the target side. The verb *dar* in Portuguese is a productive light verb, specially when combined with participles (*dar uma caminhada/corrida/passeada* lit. *to give a walk/run/stroll*). On the other hand, the translations usually involve a single verb and not a light-verb construction. This indicates that further error analysis is required, studying the three verbs in this pattern separately.

	MWE source	MWE target	# T	*ts* T	Sim	F
1	**fazer comparação**	**make comparison**	4	0.37	0.64	1.16
2	**fazer comparação**	**do comparison**	1	0.23	0.56	2.04
3	fazer comparação	make method	1	0.21	0.44	2.18
4	fazer comparação	make drug	1	0.23	0.33	2.27
1	dar início	do thing	4	0.44	0.15	1.76
2	dar início	do Sul	1	1.00	0.13	2.06
3	dar início	make vaccine	1	0.31	0.24	2.30
4	dar início	make list	1	0.26	0.24	2.37
5	dar início	make roster	1	0.24	0.21	2.47

Table 3: Pattern TAKE+N: *fazer comparação* (*make comparison*) and *dar início* (lit. *give beginning* 'to start'). Correct pairs are in bold.

6 Conclusions and Future Work

This paper constitutes our first proposal towards automatic discovery of bilingual MWE lexicons. While preliminary results are promising, the obvious next step is to design an evaluation protocol and apply it. Having this goal set, the idea is testing the approach first with other patterns and, then, making a robust evaluation.

We would also like to extrapolate this method to other language pairs and MWE categories, specially those MWE translated as single words. In this case, we are still investigating solutions but one of them consists in using monolingual word embeddings and similarity measures in order to define if the translation should be an MWE or a single word.

We believe that the method itself can be improved in many ways. For instance, we would like to design a distributional similarity measure able to focus on valid alignments. We would also like to experiment with different weights for the scores (e.g. similarity seems more important than t-score). Optimizing, that is, learning these weights from small amounts of supervised data, sounds appealing as well.

At the moment, the extraction patterns represent a bottleneck and bias the obtained results towards more plausible translations. We would like to find a way to get rid of them, specially when it comes to the target side. Another point that must be underlined is the fact that, as we are not discarding the use of word alignment in the future, we would like to perform a systematic quantitative comparison with related work and methods based on word alignment.

Acknowledgments

We would like to thank the São Paulo Research Foundation (FAPESP) for the grants #2013/50757-0 (AIM-WEST project) and #2016/13002-0 (MMeaning project). This work has been partly funded by projects PARSEME (Cost Action IC1207), PARSEME-FR (ANR-14-CERA-0001), and AIM-WEST (FAPERGS-INRIA 1706-2551/13-7).

References

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In Éric Laporte, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, pages 18–26, Beijing, China, Aug. ACL.

Wilker Aziz and Lucia Specia. 2011. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *STIL 2011*, Cuiabá, MT, Obtober.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Susanne Preuß, Kutz Arrieta, Wang Ling, Fernando Batista, and Isabel Trancoso. 2014. Linguistic evaluation of support verb constructions by openlogos and google translate. In *Proc. of the Ninth LREC (LREC 2014)*, Reykjavik, Iceland, May. ELRA.

Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, May.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proc. of the 11th Workshop on MWEs (MWE 2015)* (con, 2015), pages 19–28.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California, Jun. ACL.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. In *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing* (jou, 2010), pages 59–77.

Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In Christian Fluhr and Donald E. Walker, editors, *Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988)*, pages 609–624, Cambridge, MA, USA, Mar. CID.

Kenneth Church and Patrick Hanks. 1990. Word association norms mutual information, and lexicography. *Comp. Ling.*, 16(1):22–29.

2015. *Proc. of the 11th Workshop on MWEs (MWE 2015)*, Denver, Colorado, USA. ACL.

2010. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2), Apr.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, USA. 620 p.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proc. of the 2nd EMNLP (EMNLP-2)*, pages 97–108, Brown University, RI, USA, Aug. ACL.

Emmanuel Morin and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. In *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing* (jou, 2010), pages 79–95.

Carlos Ramisch, Laurent Besacier, and Oleksandr Kobzar. 2013. How hard is it to automatically translate phrasal verbs from English to French? In Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor, and Violeta Seretan, editors, *Proc. of the MT Summit 2013 MUMTTT workshop (MUMTTT 2013)*, pages 53–61, Nice, France, Sep.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer.

Alexandre Rondon, Helena de Medeiros Caseli, and Carlos Ramisch. 2015. Never-ending multiword expressions learning. In *Proc. of the 11th Workshop on MWEs (MWE 2015)* (con, 2015), pages 45–53.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 266–275, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014a. Detecting non-compositional mwe components using wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1792–1797, Doha, Qatar, October. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014b. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden, April. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In Chu-Ren Huang and Dan Jurafsky, editors, *Proc. of the 23rd COLING (COLING 2010) — Posters*, pages 1256–1264, Beijing, China, Aug. The Coling 2010 Organizing Committee.

Yulia Tsvetkov and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In Regina Barzilay and Mark Johnson, editors, *Proc. of the 2011 EMNLP (EMNLP 2011)*, pages 836–845, Edinburgh, Scotland, UK, Jul. ACL.

Yulia Tsvetkov and Shuly Wintner. 2014. Identification of multiword expressions by combining multiple linguistic information sources. *Comp. Ling.*, 40(2):449–468.

Sina Zarrieß and Jonas Kuhn. 2009. Exploiting translational correspondences for pattern-independent MWE identification. In Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov, and Su Nam Kim, editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 23–30, Suntec, Singapore, Aug. ACL.