# Pre-tokenization of Multi-word Expressions in Cross-lingual Word Embeddings

Naoki Otani[1] Satoru Ozaki[1] Xingyuan Zhao[1]
Yucen Li[2*] Micaelah St Johns[3*] Lori Levin[1]
[1]Carnegie Mellon University [2]Facebook [3]Stanford University
[1]{notani,sozaki,xingyuaz,levin}@andrew.cmu.edu
[2]yucenli@fb.com [3]mstjohns@stanford.edu

## Abstract

Cross-lingual word embedding (CWE) algorithms represent words in multiple languages in a unified vector space. Multi-Word Expressions (MWE) are common in every language. When training word embeddings, each component word of an MWE gets its own separate embedding, and thus, MWEs are not translated by CWEs. We propose a simple method for word translation of MWEs to and from English in ten languages: we first compile lists of MWEs in each language and then tokenize the MWEs as single tokens before training word embeddings. CWEs are trained on a word-translation task using the dictionaries that only contain single words. In order to evaluate MWE translation, we created bilingual word lists from multilingual WordNet that include single-token words and MWEs, and most importantly, include MWEs that correspond to single words in another language. We show that the pre-tokenization of MWEs as single tokens performs better than averaging the embeddings of the individual tokens of the MWE. We can translate MWEs at a top-10 precision of 30-60%. The tokenization of MWEs makes the occurrences of single words in a training corpus more sparse, but we show that it does not pose negative impacts on single-word translations.

## 1 Introduction

Cross-lingual word embeddings (CWEs) are real-valued vector representations of words in multiple languages placed in a shared vector space, with the intention that words with closer meanings have closer locations in the vector space. First, monolingual word embeddings are trained based on the hypothesis of distributional semantics (Harris, 1954) that context approximates meaning. They are learned from data in a way that words used in similar contexts have similar vectors. Following that, the monolingual word embeddings are aligned to produce CWEs. CWEs are an essential building block in modern cross-lingual methods and can also be used to induce bilingual lexicons from a small seed dictionary (Mikolov et al., 2013).

An important and overlooked fact is that before CWEs are trained, the corpus is pre-processed by a word tokenizer. This illustrates a clear limitation of the state-of-the-art CWEs: they can only align words that happen to be considered as single tokens by the word tokenizer.

Multi-word expressions (MWEs) are combinations of orthographic words, whose meaning, form, use, or distribution is non-compositional or unpredictable in some way (Sag et al., 2002; Baldwin and Kim, 2010). They come in diverse forms such as compound nouns (*dance floor*), named entities (*United States*), phrasal verbs (*give up*), and connectives (*as well as*). Word tokenizers do not recognize MWEs as single units but rather as a sequence of their components, a deficiency carried into CWE construction.

In this position paper, we argue that the token units of word embeddings should be discussed more carefully, and, in particular, that MWEs should be recognized as single units before training and evaluating word embeddings. In cross-lingual applications, MWEs are particularly important. A single token in one language is often translated into an MWE in another language. So, failure to tokenize MWEs is a critical flaw of CWEs in the task of word translation and presumably in other cross-lingual tasks as well.

Some studies (Iyyer et al., 2015; Shen et al., 2018) have suggested representing phrase and sentence embeddings by taking the average or sum of their component word vectors. However, such a simple approach is not sufficient, as the meaning

[*] This work was conducted while YL and MJ were at Carnegie Mellon University.

*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4451–4464,
November 16–20, 2020.

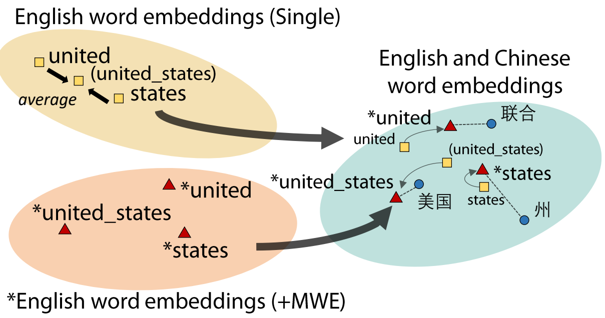


Figure 1: **The effect of MWE tokenization in cross-lingual alignments (Table 1).** English word embeddings trained with single-word tokenization (□) do not have *united states* in the vocabulary, and we represent its embedding by the average embedding. Word embeddings with MWE tokenization (△) assigns a unique embedding to *united_states*, which is better aligned with its Chinese translation 美国. Note that the configuration of single-word embeddings also changes by having MWE embeddings.

of an MWE is often unpredictable from its components, as in *red tape* and *hot dog*. Instead, MWEs should be explicitly modeled during CWE training.

To illustrate the advantage of having MWEs in the CWE vocabulary, we compare the alignments of English-Chinese CWEs with and without MWE tokens (Figure 1). Table 1 shows cosine similarities of English and Chinese words *united* and *states*. The numbers on the left side of each arrow (Single) show the cosine similarities between English and Chinese embeddings trained with standardly pre-tokenized corpora. As the English MWE *United States* is not in the vocabulary, we made an embedding for it by taking the average of the vectors of *united* and *states*. In contrast, we obtained the cosine similarities on the right-hand side of each arrow (+MWE) by combining *United States* into one token before training word embeddings.

With MWE-based tokenization, the single token *united_states* aligns with 美国 (*United States*; *meiguo*) with a high cosine similarity of 0.82. The pre-tokenization of *United States* into a single token solves additional problems as well. When we treat *United States* as two separate tokens, we distort the embeddings of *united* and *states*. On the left sides of the arrows in Table 1, both *united* and *states* have a much higher cosine similarity to 美国 than to their correct translations. Also, *united* and *state* have a higher cosine similarity to each other than they should. Recognizing *United States* as one token before training word embeddings makes it possible to translate a single token to/from an MWE and ameliorates the alignments

| Single → +MWE | 联合$_{\text{united}}$ | 州$_{\text{states}}$ | 美国$_{\text{U.S.}}$ |
|---|---|---|---|
| united | .32 → .40 | .19 → .10 | .57 → .41 |
| states | .32 → .24 | .16 → .10 | .63 → .44 |
| united_states | .37 → .38 | .20 → .18 | .69 → .82 |

Table 1: **Cosine similarities between English and Chinese word embeddings projected in the shared space.** We compare the alignments of embeddings without MWEs (left) and with MWEs (right) here.

of single tokens.[1]

In this study, we employ a simple method to identify MWEs in corpora by using MWE dictionaries instead of automatic detection. Despite the rich body of work (Constant et al., 2017), including methods developed in specialized shared tasks (Schneider et al., 2014; Savary et al., 2017; Ramisch et al., 2018), automatic MWE detection is still a hard problem (Savary et al., 2019). Ramisch et al. (2012) tested several unsupervised discovery methods and reported that they performed poorly in terms of either precision or recall.

A lexicon-based approach to MWE detection comes with another advantage. Supervised methods for MWE detection require annotated texts (Constant et al., 2017), which may not be available for all languages. On the other hand, the high availability of lexical resources containing MWEs in many languages, such as Wiktionary and WordNet, makes a lexicon-based approach for MWE detection possible in many languages.

Our focus in this paper is not to study the automatic extraction of MWEs, but rather to establish that tokenization of MWEs can contribute to improvements in CWE. Since MWE lexicons exist for the languages we are interested in, we have used those for the time being. Of course, using automatically discovered MWEs would be an interesting direction for future research.

To explore the effect of pre-tokenization of MWEs, we evaluate CWEs in the task of word translation between English and 10 languages, Arabic), Bulgarian, Chinese, German, Hebrew, Hindi, Japanese, Russian, Spanish, and Turkish, which span a wide typological variety. We find that our simple lexicon-based tokenization can align embeddings of MWEs at a precision@10 score of 30-60%

[1] The reason for the lowering of the cosine similarity between English and Chinese embeddings of *states* would be the fact that the English word *states* is polysemous while the Chinese word *states* almost exclusively means regional states. After the pre-tokenization of MWEs, English *states* no longer appears as the component of *united states*, so its distribution would be dissimilar to that of regional states.

without negative impacts on single word translation. Furthermore, we find some single-token words are correctly translated into MWEs, which are not attested in the common evaluation practice.

In summary, we argue that CWE studies should consider MWEs in development and evaluation. MWEs are pervasive in many languages and should not be ignored when the alignment of words is discussed. We present a lexicon-based method to this end (§3-4) and show its effectiveness in the task of word translation (§5). We have created a new word translation dataset that contains MWEs (§3.2). The dataset is in ten language pairs and contains MWEs in addition to single orthographic tokens.[2]

## 2 Related Work

### 2.1 Cross-lingual Word Embeddings

In this study, we experiment with one of the major approaches of learning CWEs, where monolingual embeddings trained in each language are mapped using cross-lingual supervision. Early work by Mikolov et al. (2013) showed that a linear transformation of word embeddings across languages can be trained by a bilingual dictionary. Smith et al. (2017) reported that the linear mapping becomes more accurate and computationally efficient by setting an orthogonal constraint on a transformation matrix. Recent studies (Artetxe et al., 2017; Zhang et al., 2017; Conneau et al., 2018) have further demonstrated that a transformation matrix can be learned by a very small amount of seed translations and even without any supervision.

Another stream of studies on CWEs adopts a joint approach: word embeddings on multiple languages are trained at one time using parallel corpora (Luong et al., 2015; Gouws et al., 2015). It is an interesting future direction to explore how MWEs affect joint detection of CWEs.

### 2.2 The limitations of CWEs

Besides the problem of word units, several limitations of CWEs have been pointed out in the literature. The majority of such work focuses on the statistical characteristics of word embeddings rather than their linguistic nature. Some studies (Søgaard et al., 2018; Ormazabal et al., 2019) claim that the accuracy of cross-lingual alignments depends on the similarity of word embeddings spaces of different languages, and this similarity in turn depends on the similarity between the training corpora. Kementchedjhieva et al. (2019), illustrating an issue related to evaluation of CWEs, argues that proper nouns constitute a quarter of the MUSE dataset, rendering it not ideal for word translation.

Using a word translation task for the intrinsic evaluation of CWEs presupposes a correlation between its performance with the performance of CWEs in downstream tasks, which has been questioned by several studies. Ammar et al. (2016), Glavaš et al. (2019) and Fujinuma et al. (2019) show low correlation between word translation accuracy and the performance of downstream tasks such as document classification, natural language inference, and dependency parsing. A specific problem may be that underfitting to the training data in order to better handle unseen words in the test set hinders downstream tasks that rely on words from the training dictionary (Zhang et al., 2020). In this study, we primarily examine the transferrability of MWEs in a word translation task, although it is possible that the better treatment of MWEs is also effective in downstream tasks.

### 2.3 Multi-word Expressions

MWEs have been studied in the context of syntactic analysis (Rosén et al., 2016; Kahane et al., 2017) and semantic analysis (Tratz and Hovy, 2010; Cordeiro et al., 2019). The discovery and identification of MWEs in corpora are important problems in this area (Sag et al., 2002), and much effort has been devoted to the development of methods (Constant et al., 2017) and annotated resources (Losnegaard et al., 2016). The universal dependencies (UD) project (Nivre et al., 2016) covers a wide range of languages but uses just a few dependency relations to annotate MWEs, namely *fixed*, *flat*, and *compound*. The DiMSUM shared task (Schneider et al., 2016) aims to detect English MWEs in texts. The PARSEME project (Savary et al., 2017; Ramisch et al., 2018) targets verbal MWEs and has constructed benchmark datasets in several–mostly European–languages for training automatic MWE taggers. However, such training resources are available only in a limited number of languages, and even with such resources, the automatic analysis of MWEs is known to be very difficult. Savary et al. (2019) argues the importance of syntactic MWE lexicons for further development in this area.

Another line of work analyzes the interpretation of MWEs such as noun compounds (Tratz

[2] Available at https://github.com/llab-cmu/emnlp2020-mwe-pretokenization.

and Hovy, 2010). Some studies exploit word embeddings to build a classifier (e.g., Shwartz and Waterson, 2018). Several studies tokenize MWEs before training word embeddings (Baldwin et al., 2003; Salehi et al., 2015; Cordeiro et al., 2019). Although the major target of these studies is monolingual, our focus is on the cross-lingual mapping of MWEs by CWEs.

## 3 Data Creation

This section describes the methods we used for creating the data that we are releasing with this paper: (1) monolingual lists of MWEs in eleven languages for pre-tokenizing MWEs in corpora and (2) bilingual dictionaries (ten languages each paired with English) for evaluating the resulting MWE embeddings in the word translation task. The languages are Arabic (ar), Bulgarian (bg), Chinese (zh), English (en), German (de), Hebrew (he), Hindi (hi), Japanese (ja), Russian (ru), Spanish (es), and Turkish (tr)

### 3.1 Monolingual MWE Lists for Pre-tokenization

For each of the eleven languages, we compiled a list of MWEs from publicly available resources listed below. We examined each lexical unit in each resource and selected those with multiple tokens. We treat all lexical units that are divided into two or more tokens as MWEs in our study, assuming they are fixed semantic units in some way.

**eomw:** Entries of the Extended Open Multilingual Wordnet (EOMW; Bond and Foster, 2013) consist of a WordNet synset identifier, a language identifier, and a lexical unit in that language. EOMW includes all WordNet synsets and additional synsets drawn from Wiktionary and the Unicode Common Locale Data Repository.[3] Most entries are nominals, but this resource also contains other types of MWEs like verbal phrases and connectives.

**parseme:** Parseme is multilingual corpus in which Verbal MWEs are annotated for the PARSEME shared task 1.1 (Ramisch et al., 2018). Types of verbal MWEs include light verb constructions (e.g., *give a speech*), verb-particle constructions (e.g., *wake up*), verbal idioms, etc. They can be commonly observed in many languages even though the category distributions vary from language to language.

| | eomw | | eomw | parseme |
|---|---|---|---|---|
| ar | 1,608 | bg | 1,022 | 3,255 |
| ja(i) | 5,006 | de | 1,092 | 2,705 |
| ja(u) | 3,897 | en | 8,552 | 8,982 |
| ru | 3,887 | es | 3,079 | 4,485 |
| zh | 6,927 | he | 934 | 2,454 |
| | | hi | 454 | 878 |
| | | tr | 1,959 | 4,240 |

Table 2: **MWE lists (lemma) used for MWE identification.** eomw=Extended Multilingual Open Wordnet, i=IPADIC u=UniDic

Table 2 shows the sizes of our lexicons. Note that not all MWEs in our lists are included in our word embeddings as some of them do not exist in our training corpora.

### 3.2 Bilingual Dictionaries for the Word Translation Task

Next, we built bilingual dictionaries that have MWEs for each of the pairs between English and the ten languages. To the best of our knowledge, there is no public benchmark dataset including translations between MWEs. We again used EOMW, linking lexical units in different languages with the same WordNet synset identifiers. We call the resulting bilingual dictionaries EOMW-MWE BENCHMARK, hereafter. In the EOMW-MWE benchmark, source words are all MWEs, while target words could be both single words or MWEs. We limited source words to be MWEs to ensure an MWE is always involved in translation. The number of source words varies in different language pairs. For example, zh-en has the largest number of source (zh) words, 4,813, while hi-en has 274. We report the number of source words in Table 4 (§5).

#### 3.2.1 Annotation of MWE types

We annotated the 1.5k English MWEs in our bilingual dictionaries for the purpose of error analysis.[4] We manually POS-tagged the English MWEs with the six tags **adj** (adjective phrases), **adv** (verbal and clausal adverbs), **noun** (noun phrases), **prep** (prepositional phrases), **verb** (verb phrases) and **misc** (anything else). We also classified the English

[3] We use subsets of Arabic (Elkateb and Black, 2006), Chinese (Wang and Bond, 2013), English (Fellbaum, 1998), Japanese (Isahara et al., 2008), Spanish (Gonzalez-Agirre et al., 2012), Bulgarian, Russian, German, Hebrew, Hindi, and Turkish (Bond and Foster, 2013).

[4] NO, SO, XZ and LL annotated English MWEs. LL is a professor who is a native speaker of English and has expertise in theoretical and computational linguistics. The others are non-native speakers studying computational linguistics and NLP in the US. SO also has a background in theoretical linguistics.

MWEs into four categories, **synphrase (s)**, **proper-name (pn)**, **compound (c)** and **flat+fixed+idiom (ffi)**. Below we list the definition and a prototypical example for each of the four categories.

**synphrase (s)** A semantically compositional multi-word entry from EOMW , e.g. *cease to be*.

**proper-name (pn)** A MWE that non-deictically refers to a unique or identifiable referent. Most of these are PER, LOC, GPE, or ORG in a simple NER annotation scheme. e.g. *Pacific Ocean*.

**compound (c)** We included noun-noun compounds as well as adjective-noun pairs, which are often hard to distinguish from noun-noun compounds. e.g. *opera house*, *nuclear weapon*. Most are syntactically endocentric (headed) and semantically endocentric (a hyponym of its head).

**flat+fixed+idiom (ffi)** A MWE that is one of the following: (1) A fixed grammaticalized expression that behaves like a function word or adverbial, e.g. *that is to say*; (2) A verbal idiom (e.g. *let loose*), verb-particle construction (e.g. *hang up*) or multi-verb construction (e.g. *let go*) as defined by PARSEME, and fixed collocation constructions like *take a step*, *make a decision*; (3) Any other idiomatic MWE, e.g. *bread and butter*.

We defined our own categories rather than use an existing annotation scheme. **Synphrase** was necessary because our dataset contained certain MWEs such as *other side*, *cease to be* that are frequent enough to appear in an MWE lexicon but were semantically compositional. We gave proper name its own category (**proper-name**) because proper names are uniquely nouns unlike other unheaded MWEs, which are dates, complex numerals and foreign phrases that span a wide variety of POS.

We annotated 1.5k English MWEs containing 61 s, 969 c, 215 pn, and 237 ffi. Of these 1285 are nouns, 98 verbs, 53 adjective, 52 adverb, 6 preposition, and 3 misc.[5] We excluded 18 MWEs that were numbers or contained tokenization errors.

# 4 Training CWEs: Components

This section describes our pipeline for training CWEs, including the following three steps (Figure 2): (1) identifying MWEs in a corpus, (2) training monolingual word embeddings, and (3) aligning embeddings across languages.

## 4.1 Monolingual MWE Identification

We first prepare a monolingual corpus for training word embeddings for each of the eleven languages included in this study. We take a simple lexicon-based approach to combine MWEs into one token. Suppose we have the tokenized sentence below.

(1) freedom fries was a political euphemism for french fries in the united states .

Using an MWE lexicon which includes *french fries* and *united states*, we combine tokens with underscores and obtain the following sentence.

(2) freedom fries was a political euphemism for french_fries in the united_states .

With this approach we cannot identify MWEs that do not exist in the lexicon like *freedom fries*, but there is an advantage: we do not need an annotated corpus of MWEs. Such corpora are difficult to obtain in more than a few languages.

Based on the lexicons that we compiled for each language (§3.1), we tokenize MWEs in a corpus with `mwetoolkit3` (Ramisch, 2015). To increase the recall, we use lemmas for string matching.[6] We do not consider discontinuous MWEs.

## 4.2 Monolingual Word Embeddings

We train monolingual embeddings on tokenized texts with off-the-shelf word embedding algorithms. We adopt fastText with CBOW (Bojanowski et al., 2017). MWEs processed in the previous step are treated as one token and given an individual vector. For example, *french_fries* has a different vector from those of *french* and *fries*.

## 4.3 Cross-lingual Mapping of Embeddings

Now we take two sets of word embeddings from two different languages and align the source embeddings to the target embeddings using an existing supervised method based on a bilingual dictionary. Suppose we have $n$ pairs of source and target words. We denote the embeddings of those words $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^{n\times d}$, respectively, where $d$ is the dimension of the embeddings. We learn a $d \times d$ matrix $W$ so that $XW$ is close to $Y$ in terms of Frobenius norm (Mikolov et al., 2013).

[5] Note that some MWEs have multiple possible parts-of-speech. For example, *cross over* (noun and verb).

[6] In Appendix A, we show our attempt at using unsupervised co-occurrence measures for automatic detection of MWEs for this study. We found that the vast majority of true MWEs in our evaluation lexicons had low Dice coefficient scores, which means that the automatic detection method did not predict them to have high chances of being MWEs. Thus, we were unable to find a good threshold for Dice coefficient at which precision and recall would both be adequate.

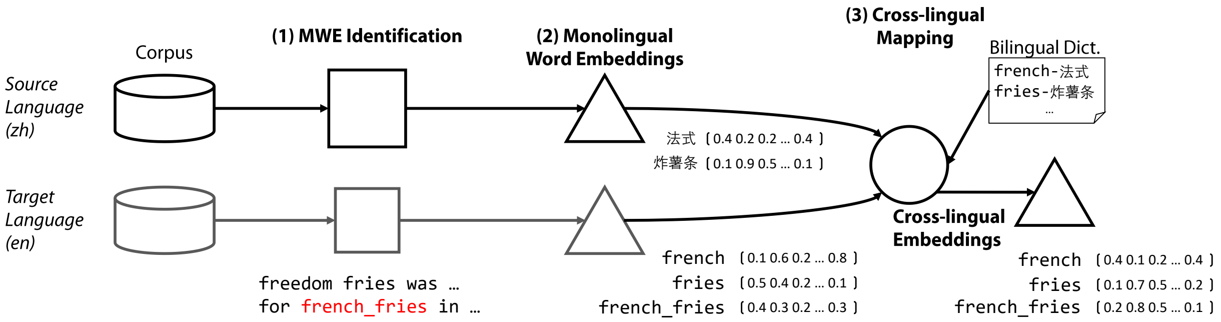

(1) MWE Identification
(2) Monolingual Word Embeddings
(3) Cross-lingual Mapping
Corpus
Source Language (zh)
Target Language (en)
Bilingual Dict.
french-法式
fries-炸薯条
…
法式 [0.4 0.2 0.2 … 0.4 ]
炸薯条 [0.1 0.9 0.5 … 0.1 ]
Cross-lingual Embeddings
freedom fries was …
for french_fries in …
french [0.1 0.6 0.2 … 0.8 ]
fries [0.5 0.4 0.2 … 0.1 ]
french_fries [0.4 0.3 0.2 … 0.3 ]
french [0.4 0.1 0.2 … 0.4 ]
fries [0.1 0.7 0.5 … 0.2 ]
french_fries [0.2 0.8 0.5 … 0.1 ]

Figure 2: **Pipeline for training CWEs with MWEs.**

$$\min_{W} |XW - Y|_{\mathrm{F}}$$

We follow Xing et al. (2015) and impose an orthogonality constraint on $W$, namely $W^T W = I$ as this constraint is known to improve the accuracy of word translation. We then refine $W$ using an iterative bootstrapping method proposed by Conneau et al. (2018). Specifically, we produce pseudo translation pairs for training by retrieving nearest neighbors in terms of cross-domain similarity local scaling (CSLS). Finally, we translate all embeddings in the source language into the vector space in the target language by $W$.

## 5 Experiments

To examine the effect of pre-tokenization of MWEs, we conduct the task of word translation between each of the ten languages and English, in both directions. A word embedding in a source language is projected into the embedding space of a target language using a trained linear mapping $W$ (§4.3). The translation candidates of the source word are retrieved by $k$-nearest neighbor search in terms of CSLS. The performance is measured by top-$k$ precision (Precision@$k$).[7]

Our evaluation involves two tasks. In the first task, we focused on the translation of MWEs using our new evaluation dictionaries that contain tokenized MWEs (§3.2). In the second task, we evaluated the translation of single words on the existing benchmark, MUSE (Conneau et al., 2018) to investigate the influence on single word embeddings of pre-tokenizing MWEs.

| Language | Sentence | Token | Type |
|---|---:|---:|---:|
| ar | 1,962,738 | 91,097,526 | 1,990,665 |
| bg | 2,739,946 | 56,871,914 | 1,643,486 |
| de | 4,961,118 | 98,123,008 | 3,439,237 |
| en | 4,174,043 | 1,00,000,031 | 1,764,082 |
| es | 3,729,100 | 99,733,231 | 1,869,469 |
| he | 3,292,840 | 84,853,134 | 1,366,709 |
| hi | 1,016,199 | 24,179,614 | 884,272 |
| ja (ipadic) | 6,709,065 | 100,000,005 | 1,164,777 |
| ja (unidic) | 3,888,640 | 100,000,004 | 2,656,774 |
| ru | 4,735,118 | 100,000,032 | 3,516,295 |
| tr | 3,055,138 | 56,576,330 | 2,011,721 |
| zh | 3,688,280 | 100,000,003 | 2,411,269 |

Table 3: **Statistics of Wikipedia corpora.**

### 5.1 Corpora

We focus on the translation between en and ten languages: ar, bg, es, de, he, hi, ja, ru, tr, and zh. These languages represent both Indo-European and non-Indo-European languages with a wide variety of morphological features and have sufficient Wikipedia texts for training embeddings. We report results using two Japanese segmentation schemes, IPADIC (Asahara and Matsumoto, 2000) and UniDic (Den et al., 2008). Both of these break Japanese utterances down into relatively small units, sometimes corresponding to morphemes. For this reason, the Japanese texts we trained on have fewer types than the other languages despite the fact that Japanese is highly agglutinative.

For training monolingual embeddings, we sampled 100M tokens for each language[8] from the publicly available Wikipedia corpora (Ginter et al., 2017), which were automatically annotated with UDPipe. Table 3 shows the corpus statistics. We then used `mwetoolkit3` to annotate MWEs. Note that the PARSEME dataset does not cover Arabic[9], Japanese, Russian, and Chinese.

[7] We used an evaluation script provided with the MUSE dictionary.

[8] Except for bg, he, hi, and tr, whose tokenized Wikipedia dumps only had 57M, 85M, 24M, and 57M tokens, respectively. We used all the texts in these languages.

[9] The PARSEME shared task covers Arabic, but the resource is not publicly available.

| en → *L2* | ar | bg | de | es | he | hi | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | 37.10 | 25.32 | 35.82 | 44.32 | 37.60 | 43.97 | 0.00 | 25.21 | 18.97 | 40.20 | 25.25 |
| MWE (eomw) | **40.23** | **37.69** | **45.71** | **56.47** | 40.60 | 45.54 | **44.76** | **40.01** | **31.97** | 44.26 | **36.29** |
| MWE (+parseme) | | 36.57 | 45.09 | 55.77 | **40.87** | **46.43** | | | | **44.68** | |
| Num. of src tokens | 1,054 | 711 | 867 | 2,279 | 734 | 448 | 1,637 | 2,217 | 2,061 | 1,184 | 1,196 |

| en ← *L2* | ar | bg | de | es | he | hi | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | 46.99 | 47.18 | 56.97 | 54.06 | 41.92 | 59.85 | 27.25 | 20.31 | 40.37 | 46.08 | 26.84 |
| MWE (eomw) | **55.22** | **54.90** | **63.11** | **64.97** | 55.39 | **65.69** | **34.77** | **29.79** | **50.19** | **53.59** | **34.20** |
| MWE (+parseme) | | **54.90** | 62.09 | 64.43 | **55.56** | 63.14 | | | | 53.43 | |
| Num. of src tokens | 1,045 | 337 | 488 | 1,687 | 594 | 274 | 3,526 | 2,481 | 1,028 | 1,211 | 4,813 |

Table 4: **Precision@10 on EOMW-MWE in Task 1.**

### 5.2 Experimental Settings

**Task 1:** In the first task, we use our EOMW-MWE dataset to evaluate the translatability of MWE embeddings obtained by lexicon-based tokenization. For some languages, the EOMW-MWE dataset has a small number of source words due to the coverage of multilingual WordNet, leaving not enough data for both training and testing. Therefore, for all languages, we used the entire EOMW-MWE dataset for testing word translation accuracy. For training, the dictionaries do not contain MWEs. The training dictionaries consist only of 5k word pairs from the common word translation benchmark, MUSE. If the cross-lingual mapping could learn a proper transformation matrix based on single word dictionaries, it should also be able to transform MWE embeddings to the shared vector space properly.

**Task 2:** We also study whether the inclusion of MWEs in cross-lingual embedding space adversely affects the alignments between single words. We use MUSE for training and evaluation in Task 2. For each language pair, we train and test cross-lingual mappings by the first 5k and next 1.5k unique source words[10] in the bilingual dictionary, respectively.

**Parameters:** We trained CBOW fastText models of 300 dimensions with the parameters suggested by Grave et al. (2018). We used the implementation by (Conneau et al., 2018) to align monolingual embeddings by the method described in §4.[11] To fairly compare between the baseline (tokenization without MWEs) and the experimental condition (tokenization with MWEs), we uses the same set of candidate words from which we are going to pick the k best. The candidate set does not include MWEs in Task 2. For the baseline in Task 1, MWEs are represented by the average of the embeddings of the individual words. We used larger vocabulary sizes (e.g, 300-600k) for the candidate set than typical sizes in related studies (e.g. 200k). We describe the details of implementation and hyperparameters in Appendix C and D.

### 5.3 Task 1: MWE Translation

As a baseline method, we tokenize the corpus without MWEs and represent the embedding of each MWE as the average of the single-word embeddings of its components. The baseline and our MWE embeddings were trained on the same single-word dictionaries. We report results of a word translation task on the EOMW-MWE in Table 4.

Despite the absence of MWEs in training dictionaries, our CWEs aligned English MWEs with their correct translation with Precision@10 as high as 30-60%. Our method clearly outperforms the baseline method in most language pairs. This fact shows the importance of learning MWE embeddings directly from a corpus to establish cross-lingual alignments.

We broke down English–*L2* MWEs translation results based on our annotated 1.5k English MWEs (§3) in Table 5. In terms of MWE types, compound (c) was the easiest category to translate (success rate of 60.22%), and flat+fixed+idiom (ffi), which includes various idiomatic expressions, was the hardest (25.52%). In terms of parts-of-speech of MWEs, it turned out that verbal MWEs were much more difficult to translate (21.01%) than nominal MWEs (48.06%). This is consistent with the observation of the PARSEME shared tasks on verbal MWE identification. Interestingly, the translation of adverbial MWEs was very accurate (40.3%). This may indicate that adverbial/adpositional phrases tend to

[10] Source words are sorted by frequencies by Conneau et al.

[11] We also experimented with VecMap (Artetxe et al., 2018) and observed a similar result (Appendix E).

| en → *L2* | | MWE type | | | | Parts-of-speech | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | s | pn | c | ffi | NOUN | VERB | ADJ | ADV | PREP | MISC |
| ar | e | 10/26 | 71/197 | 79/127 | 7/67 | 159/358 | 1/29 | 1/6 | 5/20 | 0/2 | 1/2 |
| bg | e | 2/11 | 45/142 | 50/72 | 10/57 | 96/233 | 1/16 | 1/9 | 8/21 | 1/3 | 0/0 |
| bg | +p | 4/11 | 41/142 | 48/72 | 12/57 | 92/233 | 3/16 | 1/9 | 8/21 | 1/3 | 0/0 |
| de | e | 23/49 | 354/758 | 96/164 | 48/184 | 480/1009 | 13/69 | 7/31 | 20/38 | 1/6 | 0/2 |
| de | +p | 20/49 | 363/758 | 98/164 | 54/184 | 491/1009 | 14/69 | 9/31 | 20/38 | 1/6 | 0/2 |
| es | e | 23/39 | 334/558 | 99/141 | 53/151 | 453/747 | 28/70 | 13/33 | 15/35 | 0/2 | 0/2 |
| es | +p | 22/39 | 330/558 | 99/141 | 52/151 | 451/747 | 25/70 | 12/33 | 15/35 | 0/2 | 0/2 |
| he | e | 4/16 | 57/139 | 43/73 | 11/62 | 107/235 | 0/33 | 1/7 | 7/14 | 0/1 | 0/0 |
| he | +p | 4/16 | 58/139 | 45/73 | 14/62 | 112/235 | 0/33 | 1/7 | 8/14 | 0/1 | 0/0 |
| hi | e | 1/6 | 36/68 | 39/69 | 5/25 | 77/145 | 1/10 | 0/3 | 3/9 | 0/0 | 0/1 |
| hi | +p | 1/6 | 38/71 | 40/69 | 4/25 | 79/148 | 1/9 | 0/3 | 3/10 | 0/0 | 0/1 |
| ja(i) | e | 20/46 | 242/531 | 93/161 | 36/133 | 360/763 | 16/52 | 4/19 | 11/32 | 0/3 | 0/2 |
| ja(u) | e | 18/45 | 201/503 | 83/158 | 34/132 | 305/728 | 14/53 | 5/20 | 12/32 | 0/3 | 0/2 |
| ru | e | 16/47 | 156/451 | 53/145 | 38/169 | 223/662 | 18/74 | 7/28 | 14/41 | 1/5 | 0/2 |
| tr | e | 3/22 | 132/279 | 70/103 | 16/78 | 209/417 | 1/26 | 3/14 | 7/21 | 0/2 | 1/2 |
| tr | +p | 5/22 | 131/279 | 71/103 | 17/78 | 212/417 | 1/26 | 3/14 | 7/21 | 0/2 | 1/2 |
| zh | e | 9/31 | 120/298 | 46/78 | 27/101 | 180/415 | 9/40 | 1/20 | 11/29 | 1/3 | 0/1 |
| Correct | | 38.46% | 46.14% | 60.22% | 25.52% | 48.06% | 21.01% | 24.04% | 40.3% | | |

Table 5: **Breakdown of MWE translations in Task 1.** We present two different breakdowns: (1) based on MWE categories (synphrase (s), proper-name (pn), compound (c), flat+fixed+idiom (ffi) and (2) based on parts-of-speech. The second column denotes the MWE list used for pre-tokenization: eomw (e), and eomw+parseme (+p).

| en → *L2* | ar | bg | de | es | hi | he | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | **26.21** | **35.85** | 47.97 | 64.86 | 32.00 | 28.12 | **30.37** | **31.75** | 26.09 | 31.29 | **33.62** |
| MWE (eomw) | 26.01 | 34.98 | **48.37** | **65.13** | **32.55** | **29.15** | 30.16 | 31.19 | **26.42** | 31.96 | **33.62** |
| MWE (+parseme) | | 33.98 | 46.70 | **65.13** | 32.00 | 27.36 | | | | **32.03** | |

| en ← *L2* | ar | bg | de | es | hi | he | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | **40.19** | **49.08** | **56.67** | 68.21 | **46.48** | **33.29** | 23.27 | 22.09 | 44.91 | 44.52 | 26.92 |
| MWE (eomw) | 39.16 | 48.67 | 55.20 | **68.61** | 45.00 | 33.15 | **23.81** | **22.95** | **45.52** | 44.02 | **27.58** |
| MWE (+parseme) | | 48.87 | 56.07 | 67.81 | 44.78 | 32.33 | | | | **44.73** | |

Table 6: **Precision@1 on MUSE in Task 2.**

| En | Gold | Retrieved MWE |
|---|---|---|
| chef | シェフ *shefu*<br>chef | 料理_人 *ryori_nin*<br>cooking_person |
| detect | 検出 *kenshutsu*<br>detection [n] | 検出_する *kenshutsu_suru*<br>detection_do [v] |

Table 7: **English-Japanese translation examples.**

| English MWE | Retrieved target word |
|---|---|
| in_vain | [ar] عبثا |
| that_is_to_say | [es] es_decir |
| high_school | [ja] 高校 *koko* |
| a_bit | [tr] biraz |
| dance_floor | [zh] 舞池 *wuchi* |

Table 8: **MWE translations on EOMW-MWE.**

be used in similar contexts (i.e., words in specific semantic/grammatical classes) across languages.

In Table 8, we show some correct translations retrieved by nearest neighbor search. While stop words such as "in" and "a" are usually not aligned with significant words, the inclusion of these words in MWEs (e.g., *in_vain* and *a_bit*) establishes meaningful relationships across languages.

## 5.4 Task 2: Single Word Translation

Table 6 shows the results of single-word translation on the MUSE benchmark.[12] We excluded MWEs from the embeddings in the target language as the benchmark only contains single words.

[12] We also broke down the precision scores in five language pairs based on POS of source words annotated by Kementchedjhieva et al. (2019) but did not observe meaningful patterns (Appendix E).

We were concerned that, keeping the amount of training data unchanged, the inclusion of MWEs may decrease single-word performance as it makes the occurrence of single words sparse, and it might degrade the quality of monolingual word embeddings. However, the difference in the performance of the single word translation in the other language pairs was not statistically significant.[13]

Our method might align a single word in one language with an MWE in another language, which is not attested in the common evaluation practice. To examine this, we included MWE embeddings in evaluation and observed nearest neighbors. Interestingly, our method retrieved MWEs that are correct translations but absent from the MUSE dictionaries. In particular, we show characteristic examples in English-Japanese (IPADIC) translations in Table 7. The first example illustrates a common construction using *-nin* (person), which is segmented into two words. The benchmark tends to contain transcriptions of foreign words like *shefu* as they are often single tokens. The second example shows verbalization, which is again segmented into noun + *suru* (do). These examples exemplify the limitation of evaluations restricted by single words, and may explain the difficulty of English-Japanese word translations reported in a previous study (Hoshen and Wolf, 2018).

## 6 Conclusion

We studied the impact of pre-tokenizing MWEs on cross-lingual alignments of word embeddings. We found that simple lexicon-based tokenizations can align embeddings of MWEs at a high precision without breaking alignments of single-words. We believe our results will motivate researchers to pay more attention to the existence of MWEs and how they are aligned across languages.

## Acknowledgements

We thank David R. Mortensen, Jong Hyuk Park, and anonymous reviewers for their helpful feedback.

## References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv*, abs/1602.01925.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 5012–5019. AAAI Press.

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 21–27, Saarbrücken, Germany.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan. Association for Computational Linguistics.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, Second Edition.*, pages 267–292. CRC Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *The Sixth International Conference on Learning Representations (ICLR)*.

Mathieu Constant, Gülcsen Eryiugit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

[13] We conducted pairwise bootstrapping tests with 1,000 trials for each pair of the word translation results with and without MWE tokenization. No statistical significance in our study means that the p-values are larger than 0.05.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of The Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco. European Language Resources Association.

Sabri Elkateb and William Black. 2006. Building a Wordnet for Arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC)*, pages 29–34, Genoa, Italy. European Language Resources Association.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*, volume 71. MIT Press.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul. 2019. A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4952–4962, Florence, Italy. Association for Computational Linguistics.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 710—-721, Florence, Italy. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 2525–2529, Istanbul, Turkey. European Language Resources Association.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 748–756, Lille, France. PMLR.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan. European Language Resources Association.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 469–478, Brussels, Belgium. Association for Computational Linguistics.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2420–2423, Marrakech, Morocco. European Language Resources Association.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - propositions for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 181–189, Prague, Czech Republic.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3327–3332, Hong Kong, China. Association for Computational Linguistics.

Gyri Smordal Losnegaard, Federico Sangati, Carla Parra Escartin, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Paris, France. European Language Resources Association.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv*, abs/1309.4168.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman.

2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the Limitations of Cross-lingual Word Embedding Mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4990–4995. Association for Computational Linguistics.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition*. Theory and Applications of Natural Language Processing. Springer International Publishing.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions. In *Proceedings of ACL 2012 Student Research Workshop*, pages 1–6, Jeju Island, Korea. Association for Computational Linguistics.

Victoria Rosén, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. 2016. MWEs in treebanks: From survey to guidelines. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Paris, France. European Language Resources Association.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Lecture Notes in Computer Science*, volume 2276, pages 1–15. Springer Verlag.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association for Computational Linguistics (TACL)*, 2:193–206.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–450, Melbourne, Australia. Association for Computational Linguistics.

Vered Shwartz and Chris Waterson. 2018. Olive oil is made of olives, baby oil is made for babies: Interpreting noun compounds using paraphrases in a neural model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 218–224. Association for Computational Linguistics.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *The 5th International Conference on Learning Representations (ICLR)*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Ole Tange. 2018. GNU Parallel.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 678–687, Uppsala, Sweden. Association for Computational Linguistics.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18, Nagoya, Japan. Asian Federation of Natural Language Processing.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Mozhi Zhang, Yoshinari Fujinuma, Michael J. Paul, and Jordan Boyd-Graber. 2020. Why overfitting isn't always bad: Retrofitting cross-lingual word embeddings to dictionaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2214–2220, Online. Association for Computational Linguistics.

# A Automatic MWE Discovery

In this study, we compiled MWE lists from existing lexical resources. Although MWEs can also be harvested from corpora without relying on lexical units, we found in our preliminary experiments that unsupervised methods cannot distinguish between MWEs and non-MWE phrases accurately. We tested word association measures based on word co-occurrences (Ramisch et al., 2012).

**Method:** Given tokenized texts, we extract and filter MWEs as follows:

1. We use syntactic patterns to extract candidates of MWEs. We define the following patterns based on part-of-speech (POS) tags.

   **Nominal compounds:**

   `(Adjective|Noun)+Noun`

   **Verb-particle constructions:**

   `Verb.{0,5}(up|down|on|off|in|out|away)`

   Here, $|$ denotes an OR condition, . denotes any POS tags, + denotes 1 or more repetitions, and $\{m, n\}$ denotes from $m$ to $n$ repetitions.

2. We count occurrences of MWE candidates and components of them.

3. We calculate association scores of the components of each MWE candidate by Dice coefficients, PMI, and maximum likelihood estimates (Ramisch et al., 2012).

4. We filter MWEs by setting a threshold on the association score.

Ideally, the real MWEs have higher scores, and non-MWE phrases have lower scores. Examining this, however, is not easy. It is very expensive to manually check all the candidates in Step 1. So, in our experiments, we aimed to obtain a rough estimate using the MWE lists we compiled. The phrases in our lists are true positives and should be assigned high association scores.

Figure 3 shows the results. The horizontal axis denotes the Dice coefficients calculated in Step 3, and the vertical axis shows the number of MWEs in each bin of Dice coefficients. The orange bars shows the number of MWEs that exist in our eomw+parseme lexicon, which are true positives. This result gives us two important implications.

1. The Dice coefficients are not indicative of MWE-ness. There are many true MWEs among the candidates with very low association scores. For example, the Dice coefficient of *french_fry* was only 0.000173.

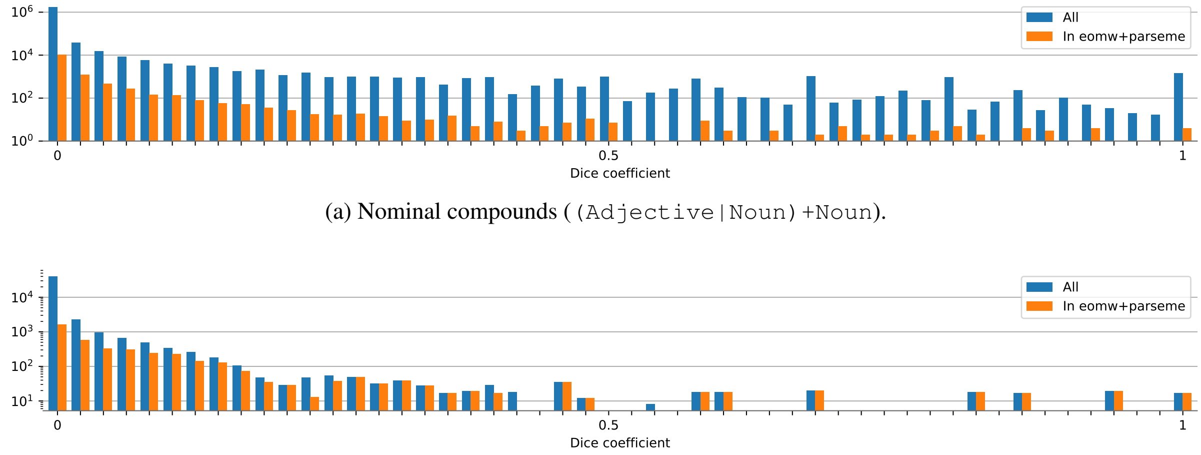
2. The distribution of the scores is highly skewed, and it is difficult to set a threshold. If we set a lower threshold, the results contain many false MWEs, and if we set a high threshold, we can only obtain a few MWEs.

We observed very similar results in association measures other than the Dice coefficients.

# B Corpus Preprocessing

We trained word embeddings on the sentences collected and torkenized following UD version 2 (Ginter et al., 2017). We lowercased texts as the tokens in MUSE dictionaries are all lowercase. The used OpenCC[14] and simplified Chinese characters. For Japanese, we tokenized plain texts provided

[14] https://github.com/BYVoid/OpenCC



(a) Nominal compounds (`(Adjective|Noun)+Noun`).

(b) Verb-particle constructions (`Verb.{0,5}(up|down|on|off|in|out|away)`).

Figure 3: Distribution of Dice coefficient scores assigned to English MWEs automatically discovered from the Wikipedia corpus. The vertical axis denotes log frequencies.

with the tokenized Wikipedia dump by MeCab with IPADIC[15]. We then sampled sentences with 100M tokens or extracted full texts. We used GNU Parallel (Tange, 2018) to speed up the preprocessing.

## C Monolingual Word Embeddings

We trained CBOW fastText models of 300 dimensions with the parameters suggested by Grave et al. (2018). Specifically, we set hyperparameters as follows:

- Dimension of word embeddings (`dim`): 300
- Minimum length of char N-gram (`minn`): 5
- Maximum length of char N-gram (`maxn`): 5
- Number of epochs (`epoch`): 10
- We set the other parameters to the default values of the fastText software v0.9.1[16].

Table 9 shows the vocabulary sizes of monolingual word embeddings. Note that the vocabulary sizes of Single are smaller than word type counts listed in Table 3 as we follow the default hyperparameters and set the minimal number of word occurrences for assigning word embeddings to 5.

## D Cross-lingual Word Embeddings

We used the supervised algorithms implemented in the MUSE library[17] and VecMap library[18] to align monolingual embeddings.

|  | Vocab. | MWE / Vocab. |  |
|---|---|---|---|
|  | **Single** | **eomw** | **+parseme** |
| ar | 374,852 | 2,296 / 376,934 |  |
| bg | 315,686 | 1,254 / 316,697 | 4,702 / 319,526 |
| de | 515,048 | 1,248 / 516,235 | 3,059 / 518,036 |
| en | 268,278 | 7,989 / 276,100 | 8,734 / 276,841 |
| es | 300,603 | 3,310 / 303,795 | 10,773 / 310,988 |
| he | 291,214 | 1,156 / 292,337 | 2,112 / 293,274 |
| hi | 124,012 | 1,147 / 125,142 | 3,344 / 127,328 |
| ja(i) | 232,299 | 4,356 / 236,579 |  |
| ja(u) | 380,605 | 3,366 / 383,777 |  |
| ru | 634,628 | 5,570 / 639,565 |  |
| tr | 350,716 | 7,122 / 357,376 | 17,451 / 367,302 |
| zh | 405,624 | 4,929 / 410,313 |  |

Table 9: **Vocabulary sizes of word embedding models.** We report the number of MWEs (the left hand side of each slash) for the MWE tokenization.

**MUSE:** We normalized word embeddings into unit vectors before training. We set the number of refinements to 1 as most of the bootstrapped word pairs were found in the first iteration.

**VecMap:** We followed the hyperparameter setting used by Artetxe et al. (2018).

## E Experimental Results

Table 10 and Table 11 show the results of Task 1 and Task 2 with supervised VecMap, respectively. The precision scores are slightly better than those of the supervised alignment with iterative refinements by Conneau et al. (2018), but the overall tendency is very similar to the result in Section 5.

Table 12 shows the result of Task 2 broken down based on the categorizations made by Kementchedjhieva et al. (2019). In some languages, the pre-tokenization of MWEs improved the translation ac-

[15] `https://taku910.github.io/mecab/`

[16] `https://github.com/facebookresearch/fastText/releases/tag/v0.9.1`

[17] `https://github.com/facebookresearch/MUSE`

[18] `https://github.com/artetxem/vecmap`

| en → *L2* | ar | bg | de | es | he | hi | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | **46.02** | 29.82 | 38.92 | 48.31 | **41.01** | **49.78** | 36.45 | 29.05 | 22.90 | **46.20** | 32.44 |
| MWE (eomw) | 43.64 | 38.12 | **46.97** | **54.80** | 40.74 | 45.54 | **45.50** | **41.68** | **33.38** | 45.95 | **37.96** |
| MWE (+parseme) | | **38.68** | 46.79 | 54.45 | 40.74 | 44.87 | | | | 45.61 | |
| Num. of src tokens | 1,054 | 711 | 867 | 2,279 | 734 | 448 | 1,637 | 2,217 | 2,061 | 1,184 | 1,196 |

| en ← *L2* | ar | bg | de | es | he | hi | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | 53.11 | 48.07 | 55.33 | 59.63 | 45.79 | 66.42 | 30.94 | 22.85 | 44.07 | 51.78 | 31.48 |
| MWE (eomw) | **57.42** | 53.12 | **61.07** | **65.26** | 55.56 | **67.88** | **37.69** | **31.24** | **53.31** | **56.73** | **38.67** |
| MWE (+parseme) | | **54.30** | 60.45 | 65.20 | **57.41** | 63.87 | | | | 56.48 | |
| Num. of src tokens | 1,045 | 337 | 488 | 1,687 | 594 | 274 | 3,526 | 2,481 | 1,028 | 1,211 | 4,813 |

Table 10: **Precision@10 of VecMap (supervised) on EOMW-MWE in Task 1.**

| en → *L2* | ar | bg | de | es | hi | he | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | **26.81** | 35.99 | 51.50 | 63.86 | **34.28** | 30.39 | 29.53 | 31.47 | 25.75 | 33.72 | **32.47** |
| MWE (eomw) | 25.74 | **37.73** | 51.23 | 65.33 | 34.00 | **30.88** | **29.95** | **32.39** | **25.89** | **34.86** | 31.04 |
| MWE (+parseme) | | 36.86 | **51.77** | **65.93** | **34.28** | 29.77 | | | | 34.46 | |

| en ← *L2* | ar | bg | de | es | hi | he | ja(i) | ja(u) | ru | tr | zh |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single | **44.46** | **50.99** | 53.99 | **70.96** | **49.00** | 34.11 | **26.73** | 25.84 | 48.82 | 47.84 | **31.02** |
| MWE (eomw) | 43.63 | 49.62 | 51.91 | 70.89 | 47.22 | **34.66** | 25.42 | **26.15** | **50.71** | **48.20** | 30.58 |
| MWE (+parseme) | | 48.60 | **55.06** | **70.96** | 46.71 | 34.11 | | | | 46.92 | |

Table 11: **Precision@1 of VecMap (supervised) on MUSE in Task 2.**

curacy of adjective, noun, and verbs (en-de (eomw), en-hi (eomw), es-en, hi-en (eomw)), but it did not in other languages. Overall, there is no clear, interpretable tendency from the results. The inclusion of MWEs in the vocabulary increased the performance of MWE translation without a negative impact on single-word translations.

To analyze the statistical significance of results, we used BOOTS[19] and conducted pairwise bootstrapping tests with 1,000 trials.

| en → *L2* | MWE | a+n+v | pn |
|---|---|---|---|
| ar | eomw | -0.37 | 0.28 |
| bg | eomw | -0.71 | -1.57 |
| | +parseme | -1.69 | -2.10 |
| de | eomw | 0.59 | -0.78 |
| | +parseme | -1.01 | -1.04 |
| es | eomw | 0.00 | 0.43 |
| | +parseme | 0.00 | 1.08 |
| hi | eomw | 0.64 | 1.39 |
| | +parseme | -0.46 | -1.11 |

| en ← *L2* | MWE | a+n+v | pn |
|---|---|---|---|
| ar | eomw | -1.10 | 0.86 |
| bg | eomw | -0.53 | -0.30 |
| | +parseme | -0.79 | 1.52 |
| de | eomw | -1.53 | -1.30 |
| | +parseme | -0.90 | 0.00 |
| es | eomw | 0.44 | 0.24 |
| | +parseme | 0.53 | -2.89 |
| hi | eomw | 1.15 | -1.53 |
| | +parseme | -0.21 | -3.36 |

Table 12: **Breakdown of the precision@1 scores on MUSE in Task 2.** Values denote the differences from the scores of the Single tokenization baselines. a=adjective, n=noun, v=verb, pn=proper noun.

[19] http://research.nii.ac.jp/ntcir/tools/boots-en.html