Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2970–2979
Marseille, 11–16 May 2020

MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora

Lifeng Han[1], **Gareth J.F. Jones**[1] **and Alan F. Smeaton**[2]
[1] ADAPT Research Centre
[2] Insight Centre for Data Analytics
School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland
lifeng.han@adaptcentre.ie; {gareth.jones, alan.smeaton}@dcu.ie

Abstract

Multi-word expressions (MWEs) are a hot topic in research in natural language processing (NLP), including topics such as MWE detection, MWE decomposition, and research investigating the exploitation of MWEs in other NLP fields such as Machine Translation. However, the availability of bilingual or multi-lingual MWE corpora is very limited. The only bilingual MWE corpora that we are aware of is from the PARSEME (PARSing and Multi-word Expressions) EU Project. This is a small collection of only 871 pairs of English-German MWEs. In this paper, we present multi-lingual and bilingual MWE corpora that we have extracted from root parallel corpora. Our collections are 3,159,226 and 143,042 bilingual MWE pairs for German-English and Chinese-English respectively after filtering. We examine the quality of these extracted bilingual MWEs in MT experiments. Our initial experiments applying MWEs in MT show improved translation performances on MWE terms in qualitative analysis and better general evaluation scores in quantitative analysis, on both German-English and Chinese-English language pairs. We follow a standard experimental pipeline to create our MultiMWE corpora which are available online. Researchers can use this free corpus for their own models or use them in a knowledge base as model features.

Keywords: Multi-lingual Corpus, Multi-word Expression, Machine Translation, Language Resource, Evaluation

1. Introduction

The use of multi-word expressions (MWEs) has become a hot topic in research in the field of natural language processing (NLP). Topics of interests in MWEs include issues such as MWE detection (Maldonado et al., 2017), MWE decomposition, and the integration of MWEs into other NLP applications such as Machine Translation (MT). However, to support research into the multilingual use of MWEs, the availability of bilingual or multi-lingual MWE corpora is very limited. The only freely available bilingual MWE corpora that we are aware of, at the submission time, is from the PARSEME (PARSing and Multi-word Expressions) EU Project[1]. This corpus is quite small, containing only 871 pairs of English-German MWEs. In this paper we present details of multi-lingual and bilingual MWE corpora that we have extracted from parallel corpora. We examine the quality of these extracted bilingual MWEs in MT experiments. Results for our initial experiments of applying MWEs into the MT process show improved translation evaluation scores on German-English and Chinese-English language pairs. These initial results justify further development of MWEs and their use in MT and potentially other NLP applications. We follow a standard experimental pipeline (Rikters and Bojar, 2017) to extract our bilingual MWEs. Our MultiMWE corpora are freely available online[2].

This paper is organized as follows: Section 2 provides some background knowledge on MWEs and MT, Section 3 lists some related works, Section 4 is the MultiMWE corpora extraction procedure, Section 5 presents some experiments on MWE integration into MT, and Section 6 is our discussion and conclusions.

2. Background

In this section, we introduce the concept of MT and MWE, and illustrate their connection with examples. This provides background to the motivation for the development of the MultiMWE corpora which forms the subject of this paper.

2.1. Machine Translation and Multiword Expressions

MT methods seek to translate one human language into another one. MT belongs to a branch of computational linguistics (CL) and artificial intelligence (AI), in which researchers try to use computational modeling to address linguistic text translation problems. It is a very challenging task for MT to achieve both accuracy of translated information and fluency at the level of a human expert's performance or what linguists expect as output. There are many reasons for this, one of which is that the use of MWEs presents a significant obstacle for a machine to learn and generate human languages in a natural form. We use three examples to illustrate the importance of correct use of MWEs in MT. We use ZH/Zh to represent Chinese, and EN/En as English. We use pinyin (pīnyīn) to annotate the Báihuà Chinese for its pronunciation and tones (phoneticism). The MT outputs in the examples were from Google Translator engine (Vaswani et al., 2017), which represented one of state-of-the-art neural models[3].

2.1.1. Example-I: Báihuà (白話)

As a first example, we show how important it is to understand Chinese expression patterns in order to express the

[1] https://typo.uni-konstanz.de/parseme/
[2] https://github.com/poethan/MWE4MT
[3] tested on 2018/10/26

ZH source:	小明去學校上課了
ZH pinyin:	Xiǎo míng qù xué xiào shàng kè le
EN reference:	Xiao Ming went to school to attend classes
EN MT output:	Xiao Ming went to school

Figure 1: Example-I Translation of Chinese (Báihuà, modern Chinese) to English. The pattern "去 (qù) ... 了 (le)" is a dis-continuous MWE often used to express the past tense "went to do something". Here it is used to express "went to somewhere (school) for something (attending classes)".

correct tense and overall information in a sentence.

Let us examine a simple plain example "小明去學校上課了 (phoneticism: xiǎo míng qù xué xiào shàng kè le)" of modern Chinese '白話 (Báihuà)', compared with ancient '文言 (Wényán)' of which we will show one example later, to English MT as in Figure 1. In this simple example, the MT output has lost the *aim* of Xiao Ming's action to go to school, i.e., what is his purpose to go there (*to attend classes*). This reflects an overall loss of adequacy. In Chinese, there is no direct past tense in the verb, so the MT needs to acquire the knowledge of language expression patterns to be able to express the tense information and purpose of the action here. The Chinese pattern "去 (qù) ... 了 (le)" is a simple **dis-continuous Chinese MWE** used to express a past tense action (went to do something, went to somewhere).

2.1.2. Example-II: Poem (詩歌)

For the second example, we will see how correct understanding of Chinese MWEs can assist disambiguation in machine learning. Conversely, the failure to understand these MWEs can lead to an incorrect translation of the ambiguous Chinese character even in very well aligned poem sentences.

The second example sentence is "年年歲歲花相似，歲歲年年人不同 (phoneticism: nián nián suì suì huā xiāng sì, suì suì nián nián rén bù tóng)" from one poem "《代悲白頭翁》" of Tang Dynasty by Xiyi Liu[4], shown in Figure 2. The source Chinese sentence is a kind of popular saying in a poetic and rhythm format with metaphor. In this example, "年年歲歲" and "歲歲年年" are aligned continuous MWEs saying 'each year'; "花" and "人" are aligned as subject 'flower' and 'human'; "相似" and "不同" are aligned as action/status '(Flowers) stay the same' while '(Humans) are changing'. For the first half of the sentence, the MT engine translated '年年歲歲 (nián nián suì suì)' into 'one year', '花 (huā)' into 'spent', and '相似 (xiāng sì)' into 'similar'. While the translation of the MWE '年年歲歲 (nián nián suì suì)' is a credible attempt since it should be 'each year', the translation of '花 (huā)' is totally wrong in this sentence since it refers to 'flower'. This is due to the ambiguity problem in language, since the Chinese character '花 (huā)' also carries a meaning of 'spend' in other situations such as in this example of Chinese Báihuà, '我花一百，你呢？ (Wǒ huā yībǎi, nǐ ne?)' means 'I spend one hundred, how about you?'. In the second half of the MT translation, 'each year is different' loses the translation of the character '人 (rén, meaning people)', i.e. *people* are different each year. This reflects an overall **loss of adequacy** which is similar to the situation of example one.

[4]劉希夷 in Chinese, 651—679, who died at early age due to this famous poem he wrote. https://zh.wikipedia.org/wiki/劉希夷

ZH source:	年年歲歲花相似，歲歲年年人不同
ZH pinyin:	Nián nián suì suì huā xiāng sì, suì suì nián nián rén bù tóng.
EN reference:	The flowers are similar each year, while people are changing year after year.
EN MT output:	One year spent similar, each year is different

Figure 2: Example-II Translation of Chinese (poem) to English. The terms "年年歲歲 (nián nián suì suì)" and "歲歲年年 (suì suì nián nián)" are continuous MWEs. "相似 (xiāng sì)" and "不同 (bù tóng)" are words with clear boundaries.

In this example, if the MT model can understand the MWEs well, i.e., "年年歲歲" aligned to "歲歲年年" and "相似" aligned to "不同", then it is easier to acquire the knowledge that "花" is aligned to "人". Since "人" is a subject here meaning "person/people", "花" should also be a noun or pronoun, which will tell the machine to translate it with higher probability into "flower (noun)" instead of "spent (verb)". We assume that the correct recognition and translation of surrounding MWEs, in general, can help the MT model to understand the sentences better overall and improve the translation of ambiguous Chinese characters.

2.1.3. Example-III: Wényán (文言)

As a third example, similar to example one (Chinese Báihuà), we show how the MT model fails to translate an ancient Chinese Wényán sentence due to the lack of **Chinese pattern expression** knowledge. Even though it is still a popular saying, the translation of this Wényán sentence is much worse than the translation of current Chinese Báihuà. This example also contains the *multi-character named entity* information as one kind of MWE.

The third example, shown in Figure 3, is a translation of the ancient Chinese 文言 (Wényán) idiom/metaphor expression to English: "燕雀安知鴻鵠之志哉？ (phoneticism: yàn què ān zhī hóng hú zhī zhì zāi?)" from the book "《史記》" [5] by Sima Qian. This Chinese expression is often used in modern language to express someone's feelings in both verbal and written format. The MWE pattern "A 安知 B 哉？" is used to express "how can A know B?" or "A does not know B". This metaphor is used to describe that some not serious or very common folks do not know the ambition or great plan of other very motivated ones.

The MT output is poor due to the model not understanding the meaning of the entities "燕雀 (yàn què, meaning

[5]From Han Dynasty, 206 BC–220 AD https://en.wikipedia.org/wiki/Records_of_the_Grand_Historian

ZH source:	燕雀安知鴻鵠之志哉?
ZH pinyin:	Yàn què ān zhī hóng hú zhī zhì zāi?
EN reference (literal):	How can a finch know the ambition of a big bird (or swan)?
EN MT output:	What is the meaning of Yanque Anzhihong?
EN reference:	The nonsense forks do not know the ambitions of the very motivated people.

Figure 3: Example-III Translation of Chinese (Wényán) to English. "A 安知 (ān zhī) something 哉 (zāi)" is a pattern to express "how can A know something" or "A does not know something". "燕雀 (yàn què)" and "鴻鵠 (hóng hú)" are named entities as one popular kind of MWE, and "之志 (zhī zhì)" is a fixed pattern expressing "*someone*'s ambition".

finch)" and "鴻鵠 (hóng hú, meaning big bird, swan)", the **fixed/patterned expressions** "安知 (ān zhī, meaning 'how to know' or 'do not know')" and "之志 (zhī zhì, meaning *someone*'s ambition)". In the MT output, we can see that it keeps 'Yanque' in the form of the original Chinese pinyin pronunciation. This may be due to the MT system not acquiring this meaning equivalent word from its training data. The MT output also failed by putting 'Anzhihong' together the pinyin pronunciation of the three Chinese characters '安知鴻', which makes no sense at all, since '安知 (ān zhī)' is one term (patterned expression) and '鴻 (hóng)' should be part of another term (named entity) '鴻鵠 (hóng hú)'. The failure to correctly interpret these kinds of expressions presents an obstacle to effective MT.

2.2. Multiword Expressions

Researchers in computational linguistics have defined MWEs in multiple ways. However, in general, these definitions agree on the following: *a MWE shall be a term including several words to express a specific concept, which is able to be decomposed, and the words combined together as an MWE are syntactically, semantically or pragmatically (some people may add 'statistically' from the computational view) idiosyncratic in nature.* (Sag et al., 2002; Baldwin and Kim, 2010; Hüning and Schlücker, 2015)

The categories of MWEs can include idioms, compound nouns, or word combinations from different kinds of Part-of-Speech (PoS) such as verb-particle or proper names. MWEs can be classified into lexical phrases and institutionalized phrases (Sag et al., 2002). Lexical phrases include fixed or semi-fixed expressions and syntactically-flexible expressions. For instance, for verb-particle structures, there are examples: pick up, give up, put on, take off, take over, etc. For idioms, there are 'you are the apple of my eye', 'kick the bucket', etc. (just to list a few). MWEs can be **continuous** or **dis-continuous** in presentation. Continuous MWEs are words grouped together without gaps, while discontinuous ones have gaps in the overall expression, e.g. some common words inserted into MWE word groups, for instance, pick *someone* up.

For noun phrase MWEs, we can find in example 2, "年年歲歲 (nián nián suì suì, noun noun noun noun)" and "歲歲年年 (suì suì nián nián, noun noun noun noun)" meaning 'each year' or 'every year'. All four Chinese characters are individually nouns, but together they can form a phrase that can be used in an adverbial function in the sentence. We can also find in example 3, "燕雀 (yàn què, noun noun)" and "鴻鵠 (hóng hú, adjective+noun)" which are noun phrases meaning finch and swan and they are also named entities.

For verbal phrase MWEs in Chinese, we can also find "安知 (ān zhī)" from example 3, meaning 'how to know / do not know'. For fixed-expressions, we can find "之志 (zhī zhì, particle+noun)" meaning "(someone)'s goal/ambition" as a noun phrase MWE.

Examples 1, 2 and 3 illustrate that it can make computational models much easier to correctly interpret the whole sentence if they can recognize continuous and discontinuous MWEs first or during model learning.

One recent book about MWE, MT and combined research including rule-based, example-based, statistical and neural MT is (Ruslan Mitkov and Seretan, 2018). This introduces MWE research focused on various kinds of MWE types and covering different languages, including English-Basque (noun+verb), French-Romanian (verb+noun collocation), named entities (Persian, Turkish, Arabic, Pashto), German nominal compounds, Dutch compound splitting, and Croatian idioms.

2.3. MWEs in MT

MWEs play a significant role in language understanding and processing tasks, including MT. This is due to their very frequent appearances and their concept specific presentation. How to recognize MWEs correctly and translate them in a meaning-preserving way, instead of merely surface word translation is a challenging task. This section introduces existing research work in this area.

MWEs in MT are related to word sense **disambiguation** (WSD) (Vickrey et al., 2005; Chan et al., 2007), phrase **boundary** detection, and semantics (Van de Cruys and Villada Moirón, 2007). Instead of a single word case in WSD, MWEs are multiple-word expressions, which can be translated in an awkward way if the translation model cannot translate the actual meaning of the MWE in the sentence and context, such as metaphorical MWEs ('apple of someone's eyes', 'kick the bucket', listed as simple examples). Addressing MWE translation also addresses the **semantic** aspects of translation in addition to issues of syntax, e.g. MWE boundary (detection) and its affects on overall sentence understanding. For instance, example 3 shows how the MT model produced very poor output due to not recognizing MWE boundaries well. Investigations into WSD have been carried out in the context of research into Neural Machine Translation (NMT). (Marvin and Koehn, 2018) shows that despite its general effectiveness NMT does not provide a full solution to the challenges of WSD. From this result, we have an indication of how challenging it is to find a solution to the issues of *multi-word sense disambiguation* is in MT. It is highlighted in (Gonzales et al., 2017) that WSD of **rare words** is especially difficult in NMT. The most recent work exploring word senses in NMT e.g. with the Transformer model includes (Tang et al., 2018).

2.3.1. SMT+MWE

We introduce research work combining SMT and MWEs here.

The earliest work that combined MT with MWEs includes (Lambert and Banchs, 2005). This applied bilingual MWE pairs to modify the word alignment procedure of MT to improve translation quality on an English-Spanish corpus. The modification function on alignment was achieved by grouping the MWEs as one token before training.

Further work includes (Ren et al., 2009) which integrated bilingual Chinese-English MWEs into the SMT toolkit Moses, (Bouamor et al., 2012) which designed models to extract continuous MWEs and integrated them into the Moses system for French-English translation, and (Skadina, 2016) which discussed various MWEs in English-Latvian MT. Recent interesting work (Ebrahim et al., 2017) focused on phrasal verb MWEs in Arabic-English phrase-based SMT. Similar to the work above, we use different bilingual MWE extraction workflows and integrate the extracted MWE pairs into training corpora.

2.3.2. NMT+MWE

This section introduces work on the incorporation of MWEs in NMT.

MWEs can appear in different kinds of examples, such as **Names Entities (NE)** (Han et al., 2013) when the entities appear as a chunk of several words. In (Li et al., 2019), the author applied a character level sequence to sequence modeling to translate named entities and then integrated this into an overall NMT system on a Chinese-to-English task. This model was originally designed to solve the unseen word translation issue, but the results show that NEs in NMT helps to improve overall translation effectiveness as measured by BLEU score. It showed the model can derive higher quality named entity alignment in the training corpus. Similarly, the work in (Ugawa et al., 2018) focuses on the difficulty of translation of compound words in the source language, by introducing an encoder for the input word at the NE tag level at each time step. Furthermore, they designed a chunk-level LSTM above the word-level one to capture the compound named entity.

In (Rikters and Bojar, 2017), the authors showed how enhancing MWEs knowledge by adding them into a corpus can improve NMT even with very simple integration. For example, they extracted bilingual MWEs in the corpus and added bilingual MWEs pairs and sentence pairs that included the MWEs into a parallel corpus to train the NMT system in English to Czech and English to Latvian MT. The authors developed an alignment visualization tool to view the improvement in MWE alignment. The neural network platform they used is from Neural Monkey (Helcl and Libovický, 2017). Our corpora construction procedure follows the pipeline designed in this work.

3. Related Work in Corpus Construction

In this section, we introduce some related work on MWE corpus construction to advance MWE research. This includes "MWE aware English Dependency corpus" from LDC[6], where they annotated English compound words in the corpus as one kind of MWE to facilitate the constituency and dependency parsing task; and the annotation of English MWEs in web reviews data (Schneider et al., 2014) [7], where they hand-annotated online review data with comprehensive MWEs including English noun, verb, and preposition super-senses (tags include communication, group, stative, location, possession, etc.). However, both these MWE corpus construction works are monolingual tasks and focus on English only.

There is some multilingual MWE corpus construction from the PARSEME research project in (Savary et al., 2018), which includes 18 European languages. However, the constructed corpus focuses on one kind of MWE (verbal MWE), is not parallel, and the size of the data varies very much from language to language (some languages have only hundreds of sentences). We built a multilingual MWE database consisting of parallel phrases that can be used to extended NLP tasks, such as translation, extend to non-European languages (e.g. Chinese), and enlarge the size into hundreds of thousands and millions of pairs.

To build our MultiMWE corpus, we used the MWE extraction pipeline from (Rikters and Bojar, 2017)[8]. We extend the extraction work into language pairs such as German-English and Chinese-English to assess the impact of MWEs on the NMT task in general and contributed the corresponding MWE extraction patterns of tested languages. Furthermore, the extracted MWEs from our experiments are freely available for MT and NLP researchers to use for their own tasks. However, the extracted MWE candidates from this framework is only the *continuous* type. In follow up work, we will design some patterns or other models to extract discontinuous MWEs, for instance, "apple of *someone's* eyes", "pick *someone* up", and "take *something* into account", etc.

4. MultiMWE Extraction Process

In this section, we present the MultiMWE corpora construction, including German-English and Chinese-English parallel MWEs extraction and give some detailed procedures.

4.1. German-English

The root parallel corpus is from the WMT2017 German-English MT training task[9]. This contains 5.8 million German-English sentences. To create a suitable bilingual MWE corpus we adopted the following procedure (Figure 4).

- Morphological tagging of De and En.
- Tagged De/En into XML format.
- Design MWE-patterns for De/En
- Extract Monolingual MWEs with MWEtoolkit

[6]https://catalog.ldc.upenn.edu/LDC2017T01
[7]http://www.cs.cmu.edu/ ark/LexSem/
[8]https://github.com/M4t1ss/MWE-Tools
[9]http://data.statmt.org/wmt17/translation-task/preprocessed/

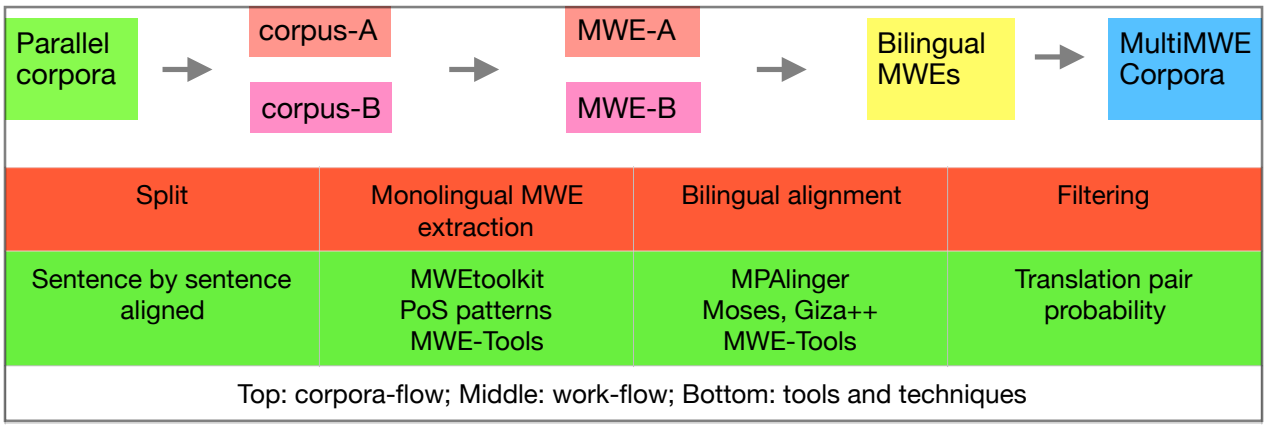

Figure 4: MultiMWE corpora extraction workflow.

- Generate De-En lexicon translation probability files with Giza++ and Moses
- Align Bilingual MWEs with MPAligner

Firstly, Treetagger[10] (Schmid, 1994) was used to tag English and German sentences with morphology information (PoS and lemmas). The English and German morphological tag-sets we used were from the BNC[11] and STTS[12] corpora. Secondly, we performed a mapping of the English POS patterns for MWE extraction from PENN used in (Rikters and Bojar, 2017) to BNC. We designed the German POS patterns for MWE extraction and then the English and German monolingual candidate MWEs were extracted using MWEtoolkit (Ramisch, 2015) with the corresponding MWE patterns and the morphological corpus. Thirdly, the MWE-tools[13] from (Rikters and Bojar, 2017) were used to convert the extracted two monolingual candidate MWE files into MPAligner format. Fourthly, we ran word alignment tools Giza++ and SMT platform Moses to get the lexical translation probability files of German-English both directions. The bilingual MWEs were aligned using MPAligner[14] (Pinnis, 2013) with the corresponding translation estimation probability. Finally, we examine the use of extracted bilingual MWEs in NMT experiments. We choose a state-of-the-art Transformer model for MT experiments, with the open package THUMT[15].

4.2. Chinese-English

To the best of our knowledge, there is no openly available bilingual MWE corpus of Chinese-English, which means we needed to create our own again. For a root parallel training corpora, we use the publicly available WMT2018 Zh-En pre-processed (word segmented) data. However, due to computational limitations in the follow-up NMT experiment, we only use the first 5 million parallel sentences as a training set.

The monolingual MWE extraction and bilingual MWE alignment procedure is mostly the same with German-English, except for some additional processing that we list as following.

- PoS pattern design.
- Stop-word list preparation.
- Zh-En translation probability files.

For the PoS patterns that MWEtoolkit required for MWE extraction, we did a PoS pattern mapping from English to Chinese and added some other Chinese PoS tags that are apparently pointing to MWEs, as shown in Table 1. These include idioms, fixed expressions, personal names, place names and organization names. The Chinese tagset is from the Lancaster Corpus of Mandarin Chinese (LCMC)[16].

In the bilingual MWE alignment step, MPaligner requires a stop word list for each language. We used some open-source packages to build a Chinese stop words file including the lists from Chinese leading IT company Baidu and NLP institutes in HIT University and Sichuan University. These packages are open source[17]. We removed duplicates when concatenating the word-lists together. There are 2,361 Chinese stop words in the merged file.

Again, we run Giza++ and Moses toolkit to get the Chinese-English lexicon translation probability files from both directions. These files will be used for bilingual MWE alignment by MPaligner.

4.3. Bilingual MWE Filtering

We manually examined the extracted bilingual MWEs and found that the MPAligner aligned bilingual MWEs have lots of noise, especially for German-English pairs. The output for German-English bilingual MWEs contains many candidates that have very low translation probabilities between 0 and 0.5. For example. The English term 'European Commission' is aligned with German 'Europäische

[10] https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[11] http://www.natcorp.ox.ac.uk/docs/c5spec.html

[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/stts_guide.pdf

[13] https://github.com/M4t1ss/MWE-Tools

[14] https://github.com/pmarcis/mp-aligner

[15] http://thumt.thunlp.org

[16] https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/

[17] https://github.com/stopwords-iso and https://github.com/goto456/stopwords

i	idiom
l	fixed expressions
nr	personal name
ns	place name
nt	orgnization name

Table 1: Added Chinese Patterns for MWEs from the LCMC Tags

German	English	Estimation
europäische Kommission	european commission	0.970964
upcoming events	european commission	0.071845
upcoming events	upcoming events	1
europäische Kommission	upcoming events	0.22279
lokaler Bürgerforen	hall meetings	0.294691
lokaler Bürgerforen	local fora	0.526792
anstehende Entscheidungen	local fora	0.131985
lokaler Bürgerforen	town hall	0.159491
lokaler Bürgerforen	town hall meetings	0.229983
anstehende Entscheidungen	town hall meetings	0.149191

Figure 5: Samples of MPAligner aligned bilingual MWEs from 5 million English-German corpora. The higher the estimation scores, the better extracted bilingual candidates.

Kommission' with 0.97 translation score, while 'upcoming events' is also aligned to German 'Europäische Kommission' with 0.22 translation probability which may be due to their co-occurrence and morphological patterns adj+noun (See Figure 5). The extracted Chinese-English bilingual MWEs generally have higher translation probability above 0.5 and are better quality (see Figure 6).

To filter out the low-quality bilingual MWE pairs, we chose two thresholds, i.e. 0.70 and 0.85, respectively in our experiments. The initially extracted and subsequently aligned bilingual MWEs are 27,688,373 pairs and 172,900 pairs for German-English and Chinese-English respectively (see discussion section for this number differences). After pruning with alignment threshold 0.70 (see Figure 7 with samples) and 0.85, the German-English MWEs moved to 6,518,550 (23.5% of original size) and 3,159,226 (11.4%) pairs. The Chinese-English MWEs moved to 143,042 pairs (82.7% of original size) with alignment threshold 0.85.

From the examples in Figure 6 and Figure 7, we can see that the extracted MWEs include some non-decomposable ones. For instance, the Chinese MWE "簸箕" (Bòji) with two characters together means "dustpan", and "電腦" (Diànnǎo) [18] means "computer". However, if we split the two characters of any of them, it can not make the same meaning. The Chinese character "電" means "electricity",

[18] we use traditional Chinese characters overall in the paper content for consistency, also to solve the character encoding issues

Chinese	English	Estimation
猫 耳	cat ears	0.780979
长 尾巴	long tail	0.820427
小 簸箕	small dustpan	0.856796
艺术 作品	artistic works	0.6281
组 表	group table	0.708438
电脑 专家	computer expert	0.801311
高尔夫球 俱乐部	golf club	0.976473
痘 产品	acne products	0.695547
不同 条件	different conditions	0.887839
常青 植物	evergreen plant	0.610852

note: (电脑)->(電腦) simplified to traditional Chinese character, used in paper content

Figure 6: Extracted Zh-En MWEs without pruning. The extracted pair samples here are from the head of the file, which have good quality.

German	English	Estimation
europäische Kommission	european commission	0.970964
upcoming events	upcoming events	1
europäischen Kommission	european commission	0.990844
praktische Informationen	practical information	0.948533
östlichen Teils	eastern part	0.793047
private Konzession	private concession	0.921197
französische Staat	french state	0.853861
europäischen Rat	european council	0.984224
größeren Infrastrukturprojekten	major infrastructure projects	0.853873
zwischengeschalteten Banken	intermediary banks	0.754617

Figure 7: Samples of Bilingual MWEs after pruning with threshold 0.70. i.e., bilingual aligned MWEs with estimation score under 0.70 are removed.

while "腦" means "brain". So the combined character sequence "電腦" is a metaphor to describe "computer". For decomposable MWEs that we extracted, there are "european commission" and "european council" as institutional names in Figure 7.

5. MWE+MT Experiments

To verify the quality of our extracted bilingual MWEs, as one example, we apply them to NMT experiments as additional knowledge to influence NMT learning. This is achieved by concatenating the extracted bilingual MWEs back to the original bilingual training corpus as additional "translation pairs". We call

the learning model with the extracted MWEs added to training corpus 'MWE+Base' and call the model with filtered MWEs "MWEpruned(threshold)+Base", e.g. MWEpruned0.7+Base.

The baseline NMT model is a state-of-the-art Transformer (THUMT-tensorflow) from (Zhang et al., 2017). This implements the all-attention based NMT encoder-decoder structure developed by Google Brain (Vaswani et al., 2017). The sub-word unit translation BPE methodology (Sennrich et al., 2016) is applied for the improvement of rare word translation. As a standard-setting, the BPE operations size is set to 32k for both German-English and Chinese-English training corpora; the vocabulary-threshold is set to 50, which means any word with frequency less than this threshold will be treated as an (out-of-vocabulary) OOV word; training set shuffling is applied by randomly relocating the order of each sentence; batch size is set at 6250. The encoder and decoder are set up with 7+7 layers.

5.1. German-English MT

The training corpus for NMT is the same as used for MWE extraction, 5.8 million parallel German-English sentences; the development and testing corpora are 3,003 and 2,169 parallel sentences respectively. To examine the external German-English MWEs that are available, we also set up one experiment where we added the 871 external MWE pairs into the training corpus. We call this ExterMWE871+Base.

After the first 20k learning steps are applied, the evaluation scores are displayed in Table 2. This result shows that even though in most n-gram matching the Baseline achieved better scores, the overall BLEU score is lower than the MWE+NMT case. This is due to the Brevity-Penalty (BP) parameter and ratio factors, which means that the MWE+NMT model produced more reference like output than the Baseline model.

It is strange that by adding 871 pairs of external De-En MWEs into the training set, the ExterMWE871+Base performance score is not higher than the baseline. The reasons could be: 1) due to the added MWEs being too small in size compared with the 5.8 million training set. 2) the external MWEs are kept as one (large) token instead of being split by the BPE model. 3) the external MWE pairs have many metaphor expressions, but such metaphors did not appear often in the training corpus, and can also mislead the learned model.

5.2. Chinese-English MT

For Chinese-English baseline NMT training, we also use the same corpora that were used for MWE extraction, 5 million parallel Zh-En sentences. The development (newsdev2017) and testing (newstest2017) corpus for NMT were from WMT2017, 2002 and 2001 parallel sentences respectively.

In the evaluation score Table 3, model MWEpruned0.85+Base means we pruned the extracted Zh-En MWEs with threshold 0.85, then we used the original BPE operators to encode the pruned MWE pairs and concatenated it to the BPE encoded training set. We used the same vocabulary file from the baseline model. The result shows that the pruned MWE pairs enhanced the model learning by producing **improved 3-gram and 4-gram BLEU scores** and yielding an overall higher score. This automatic score means that the MWE enhanced model can generally improve the chunk translation, i.e., the MT output sentences include more chunk of 3-gram and 4-grams words that match the reference sentences. Most likely, they are improved MWE translations.

When we look inside the translation outputs from the baseline model and the MWE integrated model we found some Chinese MWEs that were not translated by Baseline and were translated properly by the MWEpruned0.85+Base model. Furthermore, some idiomatic MWEs that were translated literally by Baseline, were translated in a meaning preserved way by MWEpruned0.85+Base. See Figure 8, in the first example, Chinese "口水戰" which means "war of words" was translated into "water fighting" by the Baseline, while it was translated into "oral combat" in a proper way by MWE enhanced model. The Baseline translation is due to that this is a metaphor expression in Chinese using "口水+戰" that is a combination of "saliva" and "war".

In the second example, "所謂 朋友" which means "supposed friend" is translated as "friend" in Baseline model, and this lost the Chinese MWE "所謂" which is used to express "supposed" or "so-called". The MT output yielded correct translation when we integrated the extracted bilingual MWEs back into the training corpus to enhance the learning.

However, both these two example sentences in Figure 8 showed that even though the MWE enhanced model produced better MWE translations, the BLEU scores of these two sentences do not improve correspondingly. The reason is that "oral combat" can not match reference "war of words" in the word surface form as used by BLEU metric; and "so-called" can not match reference "supposed" either.

6. Discussions and Future Work

In this work, we presented bilingual MWE corpora for German-English and Chinese-English, two typologically different languages, which we call MultiMWE-corpora. They cover 3,159,226 and 143,042 pairs of German-English and Chinese-English bilingual MWE entries after filtering. These corpora are freely available, and the size is much larger than the currently available bilingual MWE corpus. However, this current extraction procedure only generates continuous MWEs. In the future, we will design patterns to extract discontinuous MWEs or develop new extraction models.

In the current experiments, the German and Chinese PoS patterns for extracting MWEs are mapped from the English PoS tagset, via meaning equivalent alignment. In future, we plan to design German and Chinese patterns specifically for these languages and conduct some linguistic knowledge survey for this.

The NMT experiments for German-English and Chinese-English showed one example usage of the extracted bilingual MWEs, where they improved the automated translation evaluation scores slightly by BLEU metric in quantitative analysis, and assisted better MWE translations in quali-

	n-gram scores				Params		Combine
models	1-gram	2-gram	3-gram	4-gram	BP	ratio	overall
Baseline	**63.3**	**35.2**	**21.4**	**13.5**	0.942	0.944	26.73
MWEpruned0.7+Base	63.0	35.1	21.3	**13.5**	**0.952**	**0.953**	**26.87**
ExterMWE871+Base	**63.3**	**35.2**	21.2	13.3	0.929	0.932	26.15

Table 2: De-2-En NMT BLEU Scores with 20k Transformer Learning Steps.

	n-gram scores				Params		
models	1-gram	2-gram	3-gram	4-gram	BP	Ratio	overall
Baseline	**56.3**	**26.5**	14.3	8.2	**0.9**	**0.905**	18.39
MWE+Base	55.9	26.1	14.3	8.2	0.884	0.89	17.99
MWEpruned0.85+Base	55.9	26.3	**14.5**	**8.4**	0.899	0.903	**18.49**

Table 3: Zh-2-En NMT BLEU scores with 20k Transformer learning steps.

	Examples of MWE translations in MT outputs
Src	俄罗斯 与 土耳其 领导人 周二 进行 会见 ， 双方 握手 并 宣布 正式 结束 长 达 八个 月 的 口水战 与 经济制裁 。
Ref	the leaders of Russia and Turkey met on Tuesday to shake hands and declare a formal end to an eight - month long war of words and economic sanctions .
Base	Russian and Turkish leaders met Tuesday , shaking hands and declaring the official end of eight months of water fighting and economic sanctions .
B+MWE	Russian and Turkish leaders met on Tuesday and both shook hands and announced a formal end of eight months of oral combat and economic sanctions .
Src	来自 所谓 朋友 的 攻击 更让人 难以 接受
Ref	the offence was even greater , coming from a supposed friend .
Base	attacks from a friend are even harder to accept .
B+MWE	the attack from so-called friends is harder to accept .
Src: source; Ref: reference. B+MWE: Baseline+MWE. Simplified Chinese (战, 谓) mapping into Traditional (戰, 謂), used in paper.	

Figure 8: Examples of the translation outputs from Baseline model and model with filtered MWEs integrated into Baseline, with Chinese MWEs

tative analysis. By running the BLEU metric, the results are different from one language pair to another. In future work, we will explore more automated metrics that can conduct better meaning equivalent evaluation such as LEPOR(Han et al., 2012), and further investigate the translation output in more detail, such as a human in the loop evaluations and looking at MWE translations in general.

For Chinese-English MWE for NMT, we will include Chinese radicals and strokes (decomposed from Chinese characters) (Han and Kuang, 2018) into the system, and investigate the performance with these linguistic features.

When we used MWEtoolkit for Chinese monolingual MWE candidate extraction there were some issues with the toolkit for this language, which meant we had to drop out some parts of the morphologically tagged corpus. This reduced the potential MWE numbers that can be produced by this procedure. In the future, we will look at this issue and fix the toolkit for the Chinese language. This will further extend our MultiMWE corpora size for the Chinese-English pair.

We make our extracted bilingual and multilingual MWEs corpora openly available. We believe that the MultiMWE corpora can be helpful for other multilingual NLP research tasks such as multi-lingual Information Extraction (IE), Question Answering (QA), and Information Retrieval (IR). For instance, those multi-lingual / cross-lingual tasks can take MultiMWE corpora as external dictionaries/knowledge into their models.

In future work we will extend the MultiMWE corpora to other language pairs, including similar and distant languages, such as Russian-Japanese, English-French, etc. We will use the popular corpora Europarl[19] for this purpose.

7. Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. The input of Alan Smeaton is part-funded by Science Foundation Ireland under grant number SFI/12/RC/2289 (Insight Centre). We thank Matiss Rikters and Mārcis Pinnis for the supports of MWE-Tools and MPaligner, and the reviewers for valuable comments. LH thanks Paolo Bolzoni for helping experiments, Gültekin Cakir and Anna Weidmann for looking at German MWEs.

8. References

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Taylor and Francis Group.

[19] https://www.statmt.org/europarl/

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Conference on Language Resources and Evaluation*.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic*.

Ebrahim, S., Hegazy, D., Mostafa, M. G.-H. M., and El-Beltagy, S. R. (2017). Detecting and integrating multiword expression into english-arabic statistical machine translation. *Procedia Computer Science*, 117:111 – 118. Arabic Computational Linguistics.

Gonzales, A. R., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 11–19.

Han, L. and Kuang, S. (2018). Incorperating chinese radicals into neural machine translation: Deeper than character level. *Proceedings of ESSLLI-2018*, abs/1805.01565:54–65.

Han, A. L.-F., Wong, D. F., and Chao, L. S. (2012). Lepor: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 441–450. Association for Computational Linguistics.

Han, A. L.-F., Wong, D. F., and Chao, L. S. (2013). Chinese named entity recognition with conditional random fields in the light of chinese characteristics. In *Language Processing and Intelligent Information Systems*, pages 57–68. Springer.

Helcl, J. and Libovický, J. (2017). Neural Monkey: An Open-source Tool for Sequence Learning. *The Prague Bulletin of Mathematical Linguistics*, 107(1):5–17.

Hüning, M. and Schlücker, B. (2015). Multi-word expressions. *Müller et al. eds.: Word formation, An International Handbook of the Languages of Europe.*, 1:450–467.

Lambert, P. and Banchs, R. E. (2005). Data Inferred Multiword Expressions for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X*, pages 396–403, Phuket, Thailand, September.

Li, X., Yan, J., Zhang, J., and Zong, C. (2019). Neural name translation improves neural machine translation. In Jiajun Chen et al., editors, *Machine Translation*, pages 93–100, Singapore. Springer Singapore.

Maldonado, A., Han, L., Moreau, E., Alsulaimani, A., Chowdhury, K. D., Vogel, C., and Liu, Q. (2017). Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *The 13th Workshop on Multiword Expressions @ EACL 2017*. ACL.

Marvin, R. and Koehn, P. (2018). Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 125–131, Boston, MA. Association for Machine Translation in the Americas.

Pinnis, M. (2013). Context independent term mapper for European languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 562–570, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

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

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rikters, M. and Bojar, O. (2017). Paying Attention to Multi-Word Expressions in Neural Machine Translation. In *Proceedings of the 16th Machine Translation Summit (MT Summit 2017)*, Nagoya, Japan.

Ruslan Mitkov, Johanna Monti, G. C. P. and Seretan, V. (2018). *Multiword Units in Machine Translation and Translation Technology*. John Benjamins.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., and Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., and Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 455–461, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa-*

pers), pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Skadina, I. (2016). Multi-word expressions in english-latvian machine translation. *Baltic J. Modern Computing*, 4:811–825.

Tang, G., Sennrich, R., and Nivre, J. (2018). An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. *CoRR*, abs/1810.07595.

Ugawa, A., Tamura, A., Ninomiya, T., Takamura, H., and Okumura, M. (2018). Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Van de Cruys, T. and Villada Moirón, B. (2007). Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Conference on Neural Information Processing System*, pages 6000–6010.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-sense disambiguation for machine translation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 771–778, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhang, J., Ding, Y., Shen, S., Cheng, Y., Sun, M., Luan, H.-B., and Liu, Y. (2017). Thumt: An open source toolkit for neural machine translation. *ArXiv*, abs/1706.06415.