Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3103–3112
Marseille, 20-25 June 2022

Building Comparable Corpora for Assessing Multi-Word Term Alignment

Omar Adjali[1], Emmanuel Morin[2], Pierre Zweigenbaum[1]
[1]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
[2]Nantes Université, CNRS, Laboratoire des Sciences du Numérique de Nantes, France
omar.adjali@universite-paris-saclay.fr, pz@lisn.fr, emmanuel.morin@univ-nantes.fr

Abstract

Recent work has demonstrated the importance of dealing with Multi-Word Terms (MWTs) in several Natural Language Processing applications. In particular, MWTs pose serious challenges for alignment and machine translation systems because of their syntactic and semantic properties. Thus, developing algorithms that handle MWTs is becoming essential for many NLP tasks. However, the availability of bilingual and more generally multi-lingual resources is limited, especially for low-resourced languages and in specialized domains. In this paper, we propose an approach for building comparable corpora and bilingual term dictionaries that help evaluate bilingual term alignment in comparable corpora. To that aim, we exploit parallel corpora to perform automatic bilingual MWT extraction and comparable corpus construction. Parallel information helps to align bilingual MWTs and makes it easier to build comparable specialized sub-corpora. Experimental validation on an existing dataset and on manually annotated data shows the interest of the proposed methodology.

1. Introduction

The compilation of bilingual terminological resources has become critical for many NLP tasks such as cross-lingual information retrieval (Miangah, 2008), machine translation (Arcan et al., 2014; Yang et al., 2016) and many others: bilingual terminologies help such tasks either by reducing their computational cost or by improving their performance. Acquiring bilingual terminological resources is a difficult task, as it requires tremendous manual annotation effort. Early research has been conducted on automatic approaches for the acquisition of bilingual terminologies (Kupiec, 1993; Daille et al., 1994; Vintar, 2001; Wu and Chang, 2004). Bilingual term extraction (BTE) approaches first identify monolingual terms, and then establish cross-lingual correspondences between pairs of terms using alignment methods.

Two main approaches have emerged from the literature, one that relies on comparable corpora (Rapp, 1995; Tanaka and Iwasaki, 1996; Fung, 1998; Chiao and Zweigenbaum, 2002; Otero, 2007; Morin et al., 2007; Saralegi et al., 2008; Fišer et al., 2011; Fišer and Ljubešić, 2011; Ljubešic et al., 2012; Aker et al., 2013; Hazem and Morin, 2016) while the other leverages information from parallel corpora (Somers, 2001; Kwong et al., 2004; Fan et al., 2009; Lefever et al., 2009; Macken et al., 2013; Arcan et al., 2014; Yang et al., 2016; Krstev et al., 2018; Šandrih et al., 2020). Comparable corpora include non-parallel texts in different languages that share similar characteristics. While compiling them is quite easy, for example from the web, comparable corpora-based BTE needs additional external resources to reach decent performance (Otero, 2007). For example, Déjean et al. (2002) combined a multilingual thesaurus and a dictionary to extract bilingual lexicons from comparable corpora. Similarly, Nagata et al. (2001) relied on the web as a dictionary to extract English-Japanese technical terms. In contrast, parallel corpora exhibit bilingual texts that are in a translation relation. They are less widely available since building them requires considerable manual effort. Automatic methods for the collection of parallel corpora have been proposed (Resnik, 1999; Resnik and Smith, 2003; Bañón et al., 2020; Zhao and Vogel, 2002; Hangya et al., 2018), including for pairs of under-resourced languages (e.g., El-Kishky et al. (2020)). By exploiting sentence-level/document-level alignment signals, parallel corpora-based BTE approaches can better retrieve cross-lingual correspondences between pairs of bilingual terms than comparable corpora-based approaches.

Most bilingual term extraction work focused on single-word terms (SWTs). In contrast, less work addressed the bilingual extraction of multi-word terms (MWTs). Yet, some studies established that multiword terms represent the largest proportion of lexical units in a domain-specific lexicon (Constant et al., 2017). A single-/multi-word term is a type of sigle-/multi-word expression (MWE) that defines a concept from a specialized domain. Daille et al. (2004) pointed out the necessity of specifically coping with MWTs, as their inherent characteristics make MWT processing more challenging. They outlined the following properties: 1) fertility. MWTs are not always translated by a term of the same length, e.g., *"diet"* can be translated in french as *"régime alimentaire"*. 2) Like MWEs, MWTs can be characterized by the non-compositionality property i.e., the meaning of a whole MWT cannot be directly deduced by substituting each component word of a MWT by a semantically related word such as a synonym. 3) Term variation, as every MWT has different morpho-syntactic and lexical variants.

Only few bilingual MWT datasets are available (Rigouts Terryn et al., 2020), and manually annotating large datasets requires significant time and effort. Thus, we propose a methodology for automatically building a bilingual MWT dataset to evaluate MWT alignment systems. The resulting dataset offers opportunities to improve or train alignment and machine translation systems with a focus on MWTs. The proposed methodology allows to:

- Build bilingual, comparable, general-purpose corpora from parallel corpora.
- Extract a MWT bilingual terminology.
- Sample bilingual comparable specialized sub-corpora.

The proposed pipeline relies on parallel corpora to extract bilingual MWTs and align them using a bilingual embedding-based alignment approach. Parallel information is leveraged for aligning bilingual MWTs and also to build comparable corpora. We evaluated the embedding-based alignment approach on an existing dataset as well as manually annotated data. Ultimately, the same pipeline can be applied to different language pairs. The code and the dataset will be publicly available.

2. Related work

Automatic term extraction (ATE) refers to methods that output a list of potential terms in a given input specialized-domain corpus. It is worth noting that ATE can also be performed on general-purpose corpora because as stated in (Drouin et al., 2020), besides the assumption that any term can occur in general-purpose corpora, some terms related to specific topics (e.g., discrimination topic) are only included in general-purpose corpora. Automatic bilingual term extraction (BTE) requires an extra step where cross-lingual correspondences have to be established between the extracted terms in each language. This latter step can be referred to as bilingual term alignment.

Monolingual automatic term extraction

Automatic term extraction (ATE) approaches fall into three categories: Linguistic, statistical and hybrid. Linguistic approaches extract monolingual terms using symbolic methods and part-of-speech (POS) taggers. Early work such as (Dagan and Church, 1994) used regular expressions that defined syntactic patterns to match multi-word terms. Similarly, Bouamor et al. (2012) employ morphosyntactic patterns that handle both frequent and infrequent expressions without any dictionary. Savary et al. (2012) employed a graph-based method to extract polish MWTs by formulating rules that detect syntactic variation of terms, including nested terms. Similarily, Krstev et al. (2013) defined rules that handle morphological, lexical, and structural term variation.

In contrast, statistical approaches are language-independent and use various association measures to rank extracted terms. In a nutshell, word frequency and co-occurrence information are used to determine the association strength between words in a corpus. Several association measures (mutual information (MI) (Daille, 1994), C-value (Frantzi et al., 1998), T-score (Dunning, 1993) and many others) have been successfully used to rank term candidates; association-based approaches fail however to extract low-frequency terms (Pazienza et al., 2005).

Hybrid approaches take advantage of both linguistic and statistical knowledge. Daille et al. (1994) defined linguistic patterns to encode the morphosyntactic structure of MWT candidates then filtered them using statistical scores. Wu and Chang (2004) used syntactic pattern matching and cross-language statistical association measures to extract collocations from aligned sentences in a parallel corpus. A similar approach applied to the Arabic language was proposed in (Boulaknadel et al., 2008). Lefever et al. (2009) proposed a language-independent approach that is not restricted to predefined syntactic patterns, as they extract MWTs based on lexical correspondences and syntactic similarity in parallel sentences. Ranka et al. (2016) combined linguistic and statistical information using syntactic rules and association measures. The most recent approaches (Hätty and im Walde, 2018; Kucza et al., 2018; Gao and Yuan, 2019; Hazem et al., 2020) are based on deep learning models. One can find a discussion in (Rigouts Terryn et al., 2020). Among the well-established tools, hybrid methods have been evaluated as the best performing in ATE (Macken et al., 2013). In this work we rely therefore on TTC termsuite (Rocheteau and Daille, 2011; Cram and Daille, 2016).

Bilingual term alignment

Most approaches to bilingual term alignment apply monolingual ATE for each language and then perform term alignment. (DeNero and Klein, 2008; Marchand and Semmar, 2011) proposed a different strategy that considers the identification and alignment of MWTs in parallel sentences as one global problem, formulated as integer linear programming. In the present work, we focus on the more frequent strategy, which first extracts monolingual term candidates, and then applies alignment methods to detect translation correspondences.

Term alignment seeks to find correspondences between candidates across languages. Kupiec (1993) used the EM algorithm and hidden Markov Models to model term alignment. Wu and Chang (2004) extracted bilingual collocations from aligned sentences, and applied the Competitive Linking Algorithm (Melamed, 1998) to align their content words. Alternatively, following (Rapp, 1995), Chiao and Zweigenbaum (2002) used a measure of the similarity between distributional context vectors (bag of word) of source and target words to identify possible term alignments in compa-

rable corpora. Similarily, Daille et al. (2004) performed bilingual multiword term extraction following an approach based on lexical context analysis to address MWT non-compositionality and variability. Lexical context vectors are built using word co-occurrence and frequency information. Bilingual MWT association is done using a vector distance measure. Fan et al. (2009) investigated the use of statistical word aligners such as GIZA++ (Och and Ney, 2000) to extract bilingual MWTs from a Chinese-Japanese sentence-aligned corpus. Šandrih et al. (2020) also used GIZA++ to align English-Serbian MWTs. A different approach proposed by (Itagaki and Aikawa, 2008) extracts term translations using a statistical machine translation (SMT) sytem. Aker et al. (2013) formulated bilingual term extraction as a classification problem using features with a binary support vector machine classifier. Later Arcan et al. (2014) showed a more robust approach based on training a word aligner and an SMT system using parallel data in order to translate the source language terms and produce a bilingual terminology. (Morin and Daille, 2012; Liu et al., 2018b) proposed a compositional approach for the alignment of MWTs using bilingual dictionaries. In particular, (Hazem and Morin, 2017; Liu et al., 2018b) employed bilingual word embeddings for bilingual terminology extraction and showed promising results using an extended version of the bilingual word embedding mapping approach (VecMpap) of (Artetxe et al., 2016).
In this work, we follow (Liu et al., 2018b) for bilingual term alignment. This is motivated by its compositional approach that handles both SWTs and MWTs while taking advantage of advances in bilingual word embedding.

3. Bilingual term extraction method

We rely on a parallel corpus and propose a two-step approach to build a corpus for cross-lingual alignment of MWTs. Figure 1 shows the main steps of the proposed pipeline to automatically perform bilingual term extraction. We first extract monolingual MWTs from CCAligned, a large parallel corpus of aligned-sentence pairs that are translations of each other (El-Kishky et al., 2020). Given the lists of source and target MWE candidates, we align all the source-target pairs in order to find the best possible translations. We formulate the building process as bilingual extraction task in 3.1. The resulting annotated corpora will serve as resources to train and evaluate machine translation and alignment systems on MWTs.

3.1. Building task formulation

Given a pair of parallel corpora P_1 and P_2 in two different languages L_1 and L_2, the objective is to build:

- A pair of comparable corpora C_1 and C_2 in languages L_1 and L_2.
- A list of terms D_1 found in C_1 and a list of terms D_2 found in C_2.
- A reference dictionary $D_{1,2}$ in the form of a list of pairs of terms (t_1, t_2) that are translations of each other.

3.2. Parallel corpora

CCAligned is a massive dataset built from sixty-eight snapshots of the Common Crawl corpus (El-Kishky et al., 2020), where web document pairs in 8,144 language pairs, of which 137 pairs include English, have been identified such that they are translations of each other. As an example, the English-French parallel corpus contains 15,502,845 sentence-aligned pairs. They identified each document language using a text classifier (fastText), and identified pairs of cross-lingual documents using a high-precision, low-recall heuristic to assess whether two URLs represent web pages that are translations of each other. To assess their dataset construction approach, they ran a human evaluation on a diverse sample of positively-labeled documents across six language pairs.

3.3. Monolingual MWT extraction

We performed automatic term extraction (ATE) from the CCAligned parallel corpora on the English-French language pair and collected all terms, including single-word terms and multi-word terms. This provided lists of monolingual terms that represent our source and target term candidates. We used *TermSuite* (Rocheteau and Daille, 2011; Cram and Daille, 2016) for automatic term extraction. TermSuite is a multilingual terminology extractor tool that identifies term candidates using language-independent morphosyntactic patterns and ranks them according to term frequency information. It includes term a variant recognition component that improves the outputs of term extraction.
Given a source language L_1 and a target language L_2, ATE produces respectively a list of source terms T_1 and a list of target terms T_2. We filtered out single terms from each monolingual term list, keeping only MWTs. We further discarded MWTs containing proper names such as *Mr Jones*. This resulted in two MWT lists D_1 and D_2. Table 1 shows the number of terms and MWTs after monolingual ATE. We see that a great proportion of terms are MWTs.

Table 1: Extracted MWT statistics

Lang	# of Terms	# of MWTs
En	130681	48889
Fr	286581	59529

3.4. Bilingual word embedding alignment

Learning bilingual word embeddings

One approach for learning bilingual word embeddings is built on cross-lingual document-aligned/label aligned comparable corpora (Mogadala and Rettinger, 2016; Vulić and Moens, 2016; Søgaard et al., 2015),

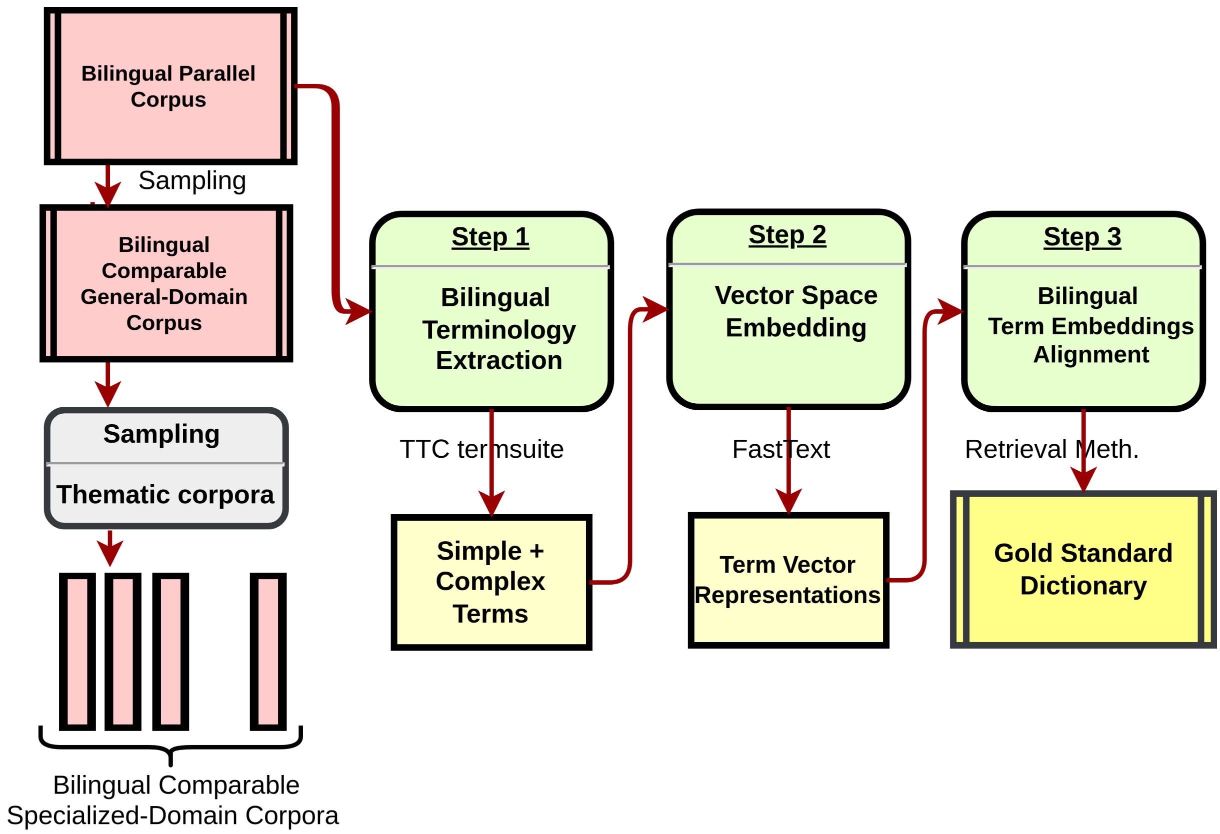

Figure 1: Bilingual MWTs extraction pipeline from parallel corpora

and parallel corpora (Gouws et al., 2015; Luong et al., 2015; Lample and Conneau, 2019). A second approach consists in mapping word representations of each language learnt separately from monolingual corpora, into a common vector space by means of linear transformations (Gaddy et al., 2016; Liu et al., 2018a; Artetxe et al., 2018). The mapping is learnt by minimizing various distances between word pairs defined in a bilingual dictionary. Hence, the mapping can alleviate the inherent limitation of dictionary-based applications such as machine translation (Artetxe et al., 2016), and computes vector representations of missing words in the dictionary. Artetxe et al. (2018) compiled a substantial number of similar methods (Mikolov et al., 2013; Faruqui and Dyer, 2014; Xing et al., 2015; Shigeto et al., 2015; Gaddy et al., 2016; Artetxe et al., 2016; Smith et al., 2017) into a multi-step bilingual word embedding framework. Although we could have followed the first bilingual word embedding approach using the CCAligned parallel corpus, we preferred the linear matrix transformation approaches as they are more time and computationally efficient.

Our method

In this work, we adopted the Compositional with Word Embedding Projection (CMWEP) approach of (Liu et al., 2018b). Note that this method handles both SWTs and MWTs with variable lenghts. It comprises the following steps:

1. Train or use pretrained monolingual word embedding models for each language to compute word vector representations. We used the 300-D fastText vectors trained on Common Crawl and Wikipedia (Bojanowski et al., 2016) and 300-D fastText vectors trained on our input parallel corpora.
2. Learn the mapping matrix following the linear transformation approach in (Artetxe et al., 2016).
3. Using a seed bilingual dictionary, compute the vector representation of each MWT in D_1 and D_2 following the compositional approach detailed in (Liu et al., 2018b). We used the English-French dictionary (113,286 entries) available in (Conneau et al., 2017).
4. For each source language MWT T_{s1}, keep as possible translation candidates only the set of MWTs $\{T_{t1},...,T_{tn}\}$ extracted from the target language sentences that are part of the bilingual sentence pairs where the source language sentences include T_{s1}. This implicitly assumes that the target term candidates are extracted during the monolingual ATE step.

5. Apply a retrieval method that helps calculate an alignment score between the vector representations of each MWT in the source list D_1 and the vector representations of their corresponding target candidatesin D_2. The candidate translations are then ranked according to their scores. This yields a first reference dictionary $D_{1,2}$.

Retrieval method

As stated in (Artetxe et al., 2018), most embedding-based bilingual lexicon extraction methods use the nearest neighbor (NN) retrieval approach: after learning the mapping matrix, for each source embedding, the closest target embedding is selected according to a similarity measure such as the cosine similarity. We extend the work of (Liu et al., 2018b) which employed the NN retrieval method and also explored more methods: 1) the inverted softmax (ISF) retrieval method (Smith et al., 2017), which replaces the cosine similarity with the softmax function while reversing the direction of the mapping query; 2) Cross-Domain Similarity Local Scaling (CSLS) (Conneau et al., 2017), which in a nutshell, computes the mean cosine similarity of each source embedding to its K target embedding neighbors (see (Conneau et al., 2017) for more details).

4. Extracting specialized comparable corpora from parallel corpora

4.1. Extracting non-parallel corpora from parallel corpora

In order to evaluate MWT alignment systems, we extract comparable corpora from the CCAligned sentence-aligned corpus. These comparable corpora will serve as bilingual resources to train and evaluate term alignment systems. Thus, given a pair of bilingual parallel corpora (C_1, C_2), we turned it into a pair of non-parallel corpora $(C_1^{'}, C_2^{'})$ by discarding one of the two sentences in each sentence pair: the L_1 sentence was discarded with probability p and the L_2 sentence with probability 1-p. Table 2 shows statistics about the constructed comparable corpora and the gold standard dictionary. We can see that the number of extracted MWTs diminished due to the sentence removal process when building the comparable corpora: MWTs that rarely occurred in the corpus have been discarded. Table 6 depicts examples from the gold standard dictionary $D_{1,2}$.

4.2. Extracting specialized sub-corpora from parallel corpora

Having a large collection made it possible to sample specialized sub-corpora on multiple topics. In a first attempt, we investigated the use of topic modeling techniques to derive specialized sub-corpora, but without satisfying results: our input parallel corpora are made of sentences, whereas topic models would probably perform better on full-document corpora. We instead used various seed lexicons found in external resources as keyword queries to select sentences and build specialized comparable sub-corpora.

4.3. Comparable medical sub-corpora

Using the Medical Subject Headings (MeSH) terminology (27,456 entries) as an input seed, we relied on the extracted lists of terms D_1 and D_2 to derive specialized comparable corpora. Following the procedure for generating non-parallel corpora, we kept only source-target non-parallel sentences that contained MeSH terms. Altogether, we extracted 340 MWT pairs included in our gold standard dictionary. Table 3 illustrates samples from the resulting medical-domain comparable corpora. The bold text shows the aligned terms from the gold standard dictionary.

Table 2: Comparable corpora: General-purpose (GPCC), Medical (MEDCC), Wind energy (WECC) English-French comparable corpora and gold standard dictionaries statistics after CC construction.

Corpus	# of sentence	# MWTs
GPCC	562030	33305
MEDCC	26904	340
WECC	3000	73

4.4. Wind energy sub-corpora

Similarly, we started with the terms of the wind energy (WE) dataset built by the TTC project (Mogadala and Rettinger, 2016). TTC released a corpus (De Groc, 2011) crawled using the Babouk crawler and a gold standard list of bilingual (En-Fr) term pairs. It is a domain-specialized corpus collected using domain-related words (wind, rotor). The gold standard list contains manually annotated En-Fr pairs: 73 MWTs and 139 SWTs.

5. Evaluation

We evaluate the quality of both steps of the proposed methodology: monolingual automatic term extraction and bilingual term alignment.

5.1. Automatic evaluation of monolingual term extraction

The evaluation of monolingual ATE is difficult and requires either to manually validate all the extracted terms, or to rely on external resources (thesaurus, dictionaries) for automatic validation. Having extracted almost 50k terms, a manual evaluation was not possible. We therefore followed the latter method, and considered valid all the terms that exist in the MeSH terminology (340 MWTs) or in the WE dataset. Obviously, the ATE tool *TermSuite* produced noise, in particular, due to the general-purpose nature of the input parallel corpora. We manually evaluated 100 sampled terms for English and French and found only respectively 5% and 8% of non valid terms. Some researchers argued

Table 3: English-French comparable corpora Examples including the multi-word term *heart disease* and its translation *maladie cardiaque*

English comparable corpus samples	French comparable corpus samples
Please inform therapist in advance if you have **heart disease**, high blood pressure or other chronic disease. hypertension pulmonaire Almost 40% of all deaths in women are related to coronary **heart disease**. Why do **heart diseases** cause so many deaths? The main contraindication, **heart disease**, a caesarean history and more than three fetal maternal disable.	Une femme sur quatre meurt d'une **maladie cardiaque** au Canada chaque année. Nous devons penser à nos grand-mères, à nos mères, à nos sœurs, à nos meilleurs amies et à nos filles. Souvenez-vous que la **maladie cardiaque** n'a pas d'âge, de race, de religion ou de penchant socio-économique. Le riz brun regorge de fibres, de lignanes et de magnésium, qui ont tous des effets bénéfiques sur la santé cardiaque et le risque de **maladie cardiaque**.

that ATE tools are specifically designed for processing specialized corpora, however, through our observational evaluation, we consider that ATE tools are viable for general-purpose corpora. Furthermore, ATE quality could be refined using methods that exploit dissimilarity between general and specialized corpora (Drouin et al., 2020).

5.2. Automatic evaluation of term alignment

We conducted experiments on the bilingual MWT alignment using the wind energy dataset. We followed the alignment procedure presented in section 3.4. We carried out two experiments, one that used fastText word embeddings pretrained on Wikipedia, and the second one employed fastText word embeddings that we trained on the input parallel corpora CCAligned. We also compared the different retrieval methods for bilingual word embeddings presented in section 3.4. We report in Tables 4 and 5 the precision (P@k) obtained by the different settings. The predicted bilingual term pairs are compared to the gold standard list of the WE dataset. Note that each source MWT has as possible target term candidates all the terms (MWTs+SWTs) present in the WE dataset.

First, we can see that the best precision scores are obtained using the fastText bilingual embeddings trained on the parallel corpora. This substantial performance boost is probably related to training on the input corpus for the task at hand and to the very large size of that input corpus. Indeed, we believe that the word embeddings carry contextual information that benefits the end task. Moreover, the results confirmed the superiority of the CSLS retrieval method whatever the embeddings. It systematically outperformed the NN and ISF retrieval methods (see Table 5), due to its ability to increase the similarity to isolated word vectors and decrease the similarity of vectors lying in dense vector spaces (Conneau et al., 2017). We also observe that the NN method performed better than ISF. This is because the ISF method needs additional hyper-parameter tuning to perform better. Finally, these results show how improvements can be obtained with the base alignment method in (Liu et al., 2018a).

Retrieval Meth.	P@1	p@5	p@10
NN	0.698	0.808	0.858
ISF	0.589	0.726	0.794
CSLS	**0.739**	0.828	0.867

Table 4: Precision of MWT alignment in the Wind Energy corpus for the language pair En-Fr using pre-trained fastText embeddings trained on Wikipedia

Retrieval Meth.	P@1	p@5	p@10
NN	0.712	0.794	0.844
ISF	0.684	0.780	0.831
CSLS	**0.780**	0.849	0.876

Table 5: Precision of MWT alignment in the Wind Energy corpus for the language pair En-Fr using fastText embeddings trained on the CCAligned Corpora.

5.3. Human evaluation of term alignment

Besides the assessment on the WE dataset, we performed a manual evaluation on the reference dictionary associated with the medical sub-corpora we extracted. We asked a native French speaker to manually validate the 340 MWTs English-French pairs. We obtained a precision (P@1) of 78.2% which demonstrates the robustness of our alignment procedure.

6. Error analysis

During the manual validation of the bilingual lexicon of the medical comparable sub-corpora, we analyzed 74 out of the 340 mis-aligned MWT pairs. Several potential sources of errors are possible, whether during the monolingual ATE, i.e., the ATE tool does not identify valid terms or identifies wrong terms, or during the alignment procedure. We will not discuss here the completeness of our gold standard dictionary, but it is very likely that a number of term pairs occur in our input corpora that are not covered by the bilingual terminology. One limitation in our parallel corpora-based BTE process lies in the inherent assumption that the monolingual ATE tool will likely extract for each source term the correct target term using the parallel aligned sentences. Obviously, this strong assumption is not always correct, as we observed that many mis-alignments occurred because the ATE failed to extract the correct target terms. For example, the extracted English medical term *"amyl nitrite"* was aligned with the French term *"médicaments à base"*. The ATE failed to extract the correct target term *"nitrate d'amyle"* that occurs in the target parallel corpus. Another limitation pointed out in (Liu et al., 2018b) is that the compositional embedding-based approach does not consider the word order in a MWT. It creates close vector representations for terms composed of the same words. The resulting ambiguity is illustrated with the following example: the English term *"water quality"* was aligned with *"eau de qualité"*. A correct alignment would have associated it with the term *"qualité de l'eau"*. Using a semantic similarity measure to align bilingual terms is also a source of errors. Among the 74 mis-aligned MWT pairs, we manually identified 50 mis-alignments where terms were not translations of each other, but close semantically. For example, the English term *"bipolar disorder"* was aligned with the term *"troubles mentaux"* instead of *"troubles bipolaires"*. Finaly, this error analysis suggests some improvements in both the monolingual ATE and alignment steps.

Table 6: English-French alignment examples

Eng term	Fr term
wind energy	énergie éolienne
third-party cookies	cookies tiers
cardiovascular disease	maladie cardiovasculaire
dark chocolate	chocolat noir
energy consumption	consommation d'énergie

7. Conclusion

In this work, we proposed a methodology for building a dataset from parallel corpora that serves as resources for evaluating bilingual MWTs alignment systems. The proposed pipeline performs bilingual MWTs extraction which results in a bilingual terminology and constructs comparable corpora. Parallel corpora are exploited for aligning bilingual MWTs and allowed to easily construct general and specialized comparable sub-corpora. Experimental validation on an existing dataset and on new manually annotated data showed the interest of the proposed methodology and also highlighted some limitations. Indeed, there is still room for improvement concerning both monolingual ATE and bilingual term alignment. In particular, our future work includes evaluating the impact of different monolingual ATE tools on the quality of the output bilingual lexicon (gold standard dictionary), and investigating cross-lingual embedding methods that exploit parallel corpora. Finally, we plan to perform BTE on several other language pairs using Multilingual CCAligned parallel corpora.

8. Bibliographical References

Aker, A., Paramita, M. L., and Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–411.

Arcan, M., Turchi, M., Tonelli, S., and Buitelaar, P. (2014). Enhancing statistical machine translation with bilingual terminology in a cat environment. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 54–68. Association for Machine Translation in the Americas.

Artetxe, M., Labaka, G., and Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Artetxe, M., Labaka, G., and Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *LREC*, pages 674–679. Citeseer.

Boulaknadel, S., Daille, B., and Aboutajdine, D. (2008). A multi-word term extraction program for arabic language. In *LREC*.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017).

Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Cram, D. and Daille, B. (2016). Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18.

Dagan, I. and Church, K. (1994). Termight: Identifying and translating technical terminology. In *Fourth Conference on Applied Natural Language Processing*, pages 34–40.

Daille, B., Gaussier, É., and Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Daille, B., Dufour-Kowalski, S., and Morin, E. (2004). French-english multi-word term alignment based on lexical context analysis. In *LREC*.

Daille, B. (1994). Study and implementation of combined techniques for automatic extraction of terminology. In *The balancing act: Combining symbolic and statistical approaches to language*.

De Groc, C. (2011). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 497–498. IEEE.

Déjean, H., Gaussier, É., and Sadat, F. (2002). Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics COLING*, pages 218–224. Citeseer.

DeNero, J. and Klein, D. (2008). The complexity of phrase alignment problems. In *Proceedings of ACL-08: HLT, Short Papers*, pages 25–28.

Drouin, P., Morel, J.-B., and L'Homme, M.-C. (2020). Automatic term extraction from newspaper corpora: Making the most of specificity and common features. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 1–7.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Fan, X., Shimizu, N., and Nakagawa, H. (2009). Automatic extraction of bilingual terms from a chinese-japanese parallel corpus. In *Proceedings of the 3rd International Universal Communication Symposium*, pages 41–45.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Fišer, D. and Ljubešić, N. (2011). Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 125–131.

Fišer, D., Ljubešić, N., Vintar, Š., and Pollak, S. (2011). Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 19–26.

Frantzi, K. T., Ananiadou, S., and Tsujii, J. (1998). The c-value/nc-value method of automatic recognition for multi-word terms. In *International conference on theory and practice of digital libraries*, pages 585–604. Springer.

Fung, P. (1998). A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 1–17. Springer.

Gaddy, D. M., Zhang, Y., Barzilay, R., and Jaakkola, T. S. (2016). Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. Association for Computational Linguistics.

Gao, Y. and Yuan, Y. (2019). Feature-less end-to-end nested term extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 607–616. Springer.

Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756. PMLR.

Hangya, V., Braune, F., Kalasouskaya, Y., and Fraser, A. (2018). Unsupervised parallel sentence extraction from comparable corpora. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT)*, pages 7–13.

Hätty, A. and im Walde, S. S. (2018). Fine-grained termhood prediction for german compound terms using neural networks. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 62–73.

Hazem, A. and Morin, E. (2016). Efficient data selection for bilingual terminology extraction from comparable corpora. In *26th International Conference on Computational Linguistics (COLING)*, pages 3401–3411.

Hazem, A. and Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 685–693.

Hazem, A., Daille, B., and Claudia, L. (2020). Towards automatic thesaurus construction and enrichment. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 62–71.

Itagaki, M. and Aikawa, T. (2008). Post-mt term swapper: Supplementing a statistical machine translation system with a user dictionary. In *LREC*.

Krstev, C., Obradović, I., Stanković, R., and Vitas, D. (2013). An approach to efficient processing of multi-word units. In *Computational Linguistics*, pages 109–129. Springer.

Krstev, C., Šandrih, B., Stanković, R., and Mladenović, M. (2018). Using english baits to catch serbian multi-word terminology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S. (2018). Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In *Interspeech*, pages 2072–2076.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 17–22.

Kwong, O. Y., Tsou, B. K., and Lai, T. B. (2004). Alignment and extraction of bilingual legal terminology from context profiles. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 10(1):81–99.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lefever, E., Macken, L., and Hoste, V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 496–504.

Liu, J., Morin, E., and Saldarriaga, S. P. (2018a). Alignement de termes de longueur variable en corpus comparables spécialisés (alignment of variable length terms in specialized comparable corpora). In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, pages 19–32.

Liu, J., Morin, E., and Saldarriaga, S. P. (2018b). Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2855–2866.

Ljubešic, N., Vintar, Š., and Fišer, D. (2012). Multi-word term extraction from comparable corpora by combining contextual and constituent clues. In *The 5th Workshop on Building and Using Comparable Corpora*, page 143.

Luong, M.-T., Pham, H., and Manning, C. D. (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.

Macken, L., Lefever, E., and Hoste, V. (2013). Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1):1–30.

Marchand, M. and Semmar, N. (2011). A hybrid multi-word terms alignment approach using word co-occurrence with a bilingual lexicon. In *Proceedings of the fifth Language and Technology Conference: Human language technologies as a challenge for computer science and linguistics*, pages 430–434.

Melamed, I. D. (1998). Word-to-word models of translational equivalence. *arXiv preprint cmp-lg/9805006*.

Miangah, T. M. (2008). Automatic term extraction for cross-language information retrieval using a bilingual parallel corpus. In *Proceedings of the 6th International Conference on Informatics and Systems (INFOS2008)*, pages 81–84. Citeseer.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mogadala, A. and Rettinger, A. (2016). Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702.

Morin, E. and Daille, B. (2012). Revising the compositional method for terminology acquisition from comparable corpora. In *Proceedings of COLING 2012*, pages 1797–1810.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining-using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671.

Nagata, M., Saito, T., and Suzuki, K. (2001). Using the web as a bilingual dictionary. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 440–447.

Otero, P. G. (2007). Learning bilingual lexicons from comparable english and spanish corpora. *Proceedings of MT Summit xI*, pages 191–198.

Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge mining*, pages 255–279. Springer.

Ranka, S., Cvetana, K., Ivan, O., Biljana, L., and Aleksandra, T. (2016). Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the 10th International Conference on Language*

Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016, pages 507–514.

Rapp, R. (1995). Identifying word translations in non-parallel texts. pages 320–322, June.

Resnik, P. and Smith, N. (2003). The Web as a parallel corpus. *Computational Linguistics*, 29:349–380. Special Issue on the Web as a Corpus.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA, June. Association for Computational Linguistics.

Rigouts Terryn, A., Hoste, V., Drouin, P., and Lefever, E. (2020). Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 85–94. European Language Resources Association (ELRA).

Rocheteau, J. and Daille, B. (2011). TTC TermSuite - a UIMA application for multilingual terminology extraction from comparable corpora. In *Proceedings of the IJCNLP 2011 System Demonstrations*, pages 9–12, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Šandrih, B., Krstev, C., and Stanković, R. (2020). Two approaches to compilation of bilingual multi-word terminology lists from lexical resources. *Natural Language Engineering*, 26(4):455–479.

Saralegi, X., San Vicente, I., and Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of Building and using Comparable Corpora workshop*, pages 27–32.

Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A., and Makowiecki, F. (2012). Sejfek-a lexicon and a shallow grammar of polish economic multi-word units. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 195–214.

Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., and Matsumoto, Y. (2015). Ridge regression, hubness, and zero-shot learning. In Annalisa Appice, et al., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 135–151, Cham. Springer International Publishing.

Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Søgaard, A., Agić, Ž., Alonso, H. M., Plank, B., Bohnet, B., and Johannsen, A. (2015). Inverted indexing for cross-lingual nlp. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1713–1722.

Somers, H. (2001). Bilingual parallel corpora and language engineering. In *Anglo Indian Workshop "Language Engineering for South Asian Languages" LESAL*.

Tanaka, K. and Iwasaki, H. (1996). Extraction of lexical translations from non-aligned corpora. In *COLING*, pages 580–585. Citeseer.

Vintar, S. (2001). Using parallel corpora for translation-oriented term extraction. *Babel*, 47(2):121–132.

Vulić, I. and Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Wu, C.-C. and Chang, J. S. (2004). Bilingual collocation extraction based on syntactic and statistical analyses. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 1, February 2004: Special Issue on Selected Papers from ROCLING XV*, pages 1–20.

Xing, C., Wang, D., Liu, C., and Lin, Y. (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*, pages 1006–1011.

Yang, W., Yan, J., and Lepage, Y. (2016). Extraction of bilingual technical terms for chinese-japanese patent translation. In *Proceedings of the NAACL Student Research Workshop*, pages 81–87.

Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 745–748.