Word Representations, Seed Lexicons, Mapping Procedures, and Reference Lists: What Matters in Bilingual Lexicon Induction from Comparable Corpora?*

Martin Laville[1], Mérième Bouhandi[1], Emmanuel Morin[1], and Philippe Langlais[2]

[1] LS2N, UMR CNRS 6004, Universite de Nantes, France
firstname.lastname@ls2n.fr
[2] Université de Montréal, Montreal, Quebec H3C 3J7, Canada
felipe@iro.umontreal.ca

Abstract. Methods for bilingual lexicon induction are often based on word embeddings (WE) similarity. These methods must be able to project the WE to the same space. Uncontextualized WE proved to be useful for this task. We compare them to contextualized WE and Bag of Words, using specialized and general datasets. We also evaluate the impact of seed lexicons and check the existing reference lists validity, claiming that extracting the translation of some words in those lists is not useful and confirming the need to have more fine-grained reference lists.

1 Introduction

Bilingual lexicons are mainly made of word pairs considered to be word-level translations of each other. They are an essential resource for several bilingual tasks, such as machine translation, cross-lingual information retrieval, and their automatic extraction, from parallel and comparable corpora, is a very active research topic. With word embeddings (WE) [3,7], being greatly in fashion these past few years and with the emergence of various mapping methods to project different languages in the same embedding space [2,8], several solutions to compare word meaning across languages have been implemented.

The recent surge of contextual embedding models [5,9] allows an interesting extension of previous work on uncontextualized WE, and various solutions have been built [11,12] to adapt these WE to actual mapping methods.

In this work, we challenge the current evaluation protocol by studying the reference lists used for bilingual lexicon induction (BLI) from comparable corpora. These lists are often used as-is, and there is a general tendency to think that the larger the list size, the more significant the results, even if they are filled with proper names, perfect cognates or even incorrect words in the language of interest. We examine these issues by filtering down general and specialized reference lists into sublists, examining the resulting differences when using supervised and unsupervised methods and the Bag of Words (BoW) method.

*This research has received funding from the French National Research Agency under grant ANR-17-CE23-0001 and the Canadian Institute for Data Valorisation.

Our main contributions seek to observe and understand the difference between BLI techniques when using specialized and general corpora and to take a more critical and precise look at these lists.

2 Methods

We compare three word representations (BoW, uncontextualized and contextualized WE methods) and various mapping methods (unsupervised and supervised).

2.1 Bag of Words

The Distributional Standard Approach [10] is the historical method for BLI from comparable corpora. Based on the idea that a word is defined by its context, the semantic proximity of two words is determined by the degree of overlap of their contexts. For each source language word, its context vector is translated into the target language using a bilingual seed lexicon, allowing the source word and its translation to appear in similar contexts, enabling their alignment.

For specialized corpora, [6] shows that adding general data to the corpus improves the representations of general words, allowing for an increase in results.

2.2 Embedding Methods

The introduction of deep distributed representations [7] renewed this historical method. In [8] the authors proposed an approach to learn a linear transformation from the source to the target embedding spaces.

We use fastText [3] as our uncontextualized WE. For contextualized WE, we use ELMo [9] and pre-trained models from [11]. To make contextualized embeddings suitable for classic mapping methods, we follow [11], creating anchors for each word by averaging the embeddings of each occurrences of this word.

After extracting our embeddings separately from the source and target corpora, and in order to be able to compare them, we map the obtained matrices into the same space.

We use two different approaches to map both fastText and anchored ELMo embeddings. The unsupervised one [2] creates and refines an initial seed lexicon based on the idea that source (X) and target (Y) embeddings space are perfectly isometric. The similarity matrices ($M_X = XX^T$ and $M_Y = YY^T$) are, then, just a permutation of their rows and columns. We also experiment with a supervised [1] method, with seed lexicon of different sizes.

2.3 Cross-domain Similarity Local Scaling (CSLS)

After mapping the word vectors in a shared space, we measure the similarity between each source word of the reference list and the target words. Since the usage of cosine similarity measure suffers from the hubness problem (some points tend to be nearest neighbours to many others), we reduce the similarity for word vectors in dense areas and increase it for isolated ones by using CSLS [4].

3 Data and analysis

In this section, we describe the corpora, seed lexicons and reference lists used in most evaluations conducted recently [11,12].

We use two English/French comparable corpora. The Breast Cancer (BC) corpus represents our specialized domain and contains 500,000 words for each language. It is composed of scientific documents, available in open access on the ScienceDirect portal, where the title or the keywords contain the term *breast cancer* in English (and their French translation). Our general corpus is a fraction of the same Wikipedia (WIKI) dumps than [11] (100M words for each language).

For the supervised mapping and the BoW approach, we use a first seed lexicon (10,872 pairs) from MUSE [4] and a second one from the general domain ELRA-M0033 French/English dictionary (243,539 pairs).

As for evaluation (see Table 1), we use one general domain reference list from [4], and a specialized one from the UMLS. The reference list in the specialized domain is smaller than the general one because it is harder to find many words in the same specialized domain if we do not want to incorporate less specific words. However, the specialized reference list represents a more significant part of its corpus vocabulary than the general one does (6% versus 0.5%).

Domain	Original	In-dictionary	Lev. ≥ 3	Freq. ≤ 100
General	1,446	1,139 (79%)	783 (54%)	146 (10%)
Specialized	248	216 (87%)	85 (34%)	18 (8%)

Table 1: Size of the reference lists and their sublists

In the general domain list, we found many words that do not belong to the language of interest (i.e. *garrison* or *enjoy* being in the French part). Translating such entities is not of much interest and pollutes the conducted evaluation. In order to verify this claim, we filter our lists with monolingual general dictionaries[3,4], removing also proper names in the process, but isolating a subset of pairs that makes more sense to translate. This cuts down the general domain list by almost a quarter of its words. We also apply the dictionary filtering on the specialized domain reference. Here, unlike for the general domain, we see that most of the words found to be out-of-dictionary are acronyms (e.g., *DNA*, *AIDS*) or in-domain words which are not part of a standard dictionary that we still want to translate because they are part of the specialized domain.

Furthermore, in the remaining pairs for both general and specialized lists, we found many pairs with nearly identical words. Even if usage of monolingual dictionaries already solved part of this with the deletion of proper names and

[3]English dictionary: `github.com/dwyl/english-words`

[4]French dictionary: `infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html`

city or country names, we use the Levenshtein distance (the number of deletions, insertions, or substitutions required to transform a string to another) and add a third reference list trying to study words with no shared morphology.

We create one last sublist to study rare words, only keeping the ones appearing less than 100 times, dropping the size of our specialized domain list to only 18 pairs, making it quite hard to draw meaningful conclusions.

4 Results

Table 2 shows the results obtained using the different reference lists, word representations models, and mapping approaches on the general and specialized domains. For the general domain, we extract the vectors from WIKI. For the specialized domain, we enrich BC with general data from WIKI.

We can observe that, even if the results of the unsupervised methods are close to the supervised ones, the latter is still the way to go, being at least two points higher in most configurations. We can also see that using a more substantial dictionary does not mean getting better results for ELMo and fastText.

Domain	Mapping	Embeddings	Original	In-dictionary	Lev. ≥ 3	Freq. ≤ 100
General	Unsupervised	fastText	**68.9**	60.2	38.4	30.8
		ELMo	62.1	**72.2**	**57.2**	**68.0**
	Supervised (MUSE)	fastText	**70.4**	64.6	44.1	44.9
		ELMo	63.4	**72.7**	**59.0**	**70.1**
		BoW	53.4	49.9	35.7	4.3
	Supervised (ELRA)	fastText	63.8	63.2	44.2	41.7
		ELMo	57.4	70.1	55.9	58.5
		BoW	43.8	46.2	34.7	3.8
Specialized	Unsupervised	fastText	**80.6**	**81.4**	60.0	**94.4**
		ELMo	70.4	77.7	**61.2**	61.1
	Supervised (MUSE)	fastText	**81.8**	**82.3**	**63.5**	**83.3**
		ELMo	68.4	75.3	62.4	50.0
		BoW	59.5	65.3	53.8	16.7
	Supervised (ELRA)	fastText	80.2	81.9	62.4	77.8
		ELMo	68.8	75.8	61.2	50.0
		BoW	67.6	73.5	61.2	27.8

Table 2: P@1 (%) on multiple lists using different word representations.

The results obtained with the three mapping methods all have the same trends for the different lists, but the list-based variations are more significant.

On the original list, fastText always gets the most interesting results. However, when we filter the list down, its results degrade notably, while ELMo is way less affected, showing that fastText is better at predicting graphically close words since it works with character n-grams. BoW is left behind, especially with the filtered lists (less than 5% for general domain and with frequency ≤ 100).

5 Analysis

In this section, we provide a more qualitative analysis of the results obtained for both general and specialized domains to illustrate the trends mentioned above from studied lists. To do this, we show in Table 3 some word pairs with their frequencies and their n-best translations as found by the different approaches.

For the general domain, we observe that fastText mostly finds graphically close words, without really grasping the concept behind the words ("napoléone" is a plant, and "rings" while "wrestlers" are not French words). Conversely, ELMo seems to capture their meaning, finding war-related concepts for "napoleon", or geometric shapes for "rings". BoW seems really affected by word occurrences.

In the specialized domain, since words are supposed to have only one specific meaning, they are less likely to be found in varying contexts. FastText and BoW do a better job at understanding these words, even if they can have some problems for infrequent words like "vincristine".

Domain	Method	Word *Translation*	Top 1	Top 2	Top 3	Top 4
General	fastText	napoleon: 2.1k *napoléon*: 5.2k	napoléon	napoléone	napoléonienne	napoléonnien
	ELMo		bélisaire	napoléon	guerry	salinator
	BoW		napoléon	bonaparte	xiv	prussien
	fastText	rings: 710 *anneaux*: 117	anneaux	rings	ring	anneau
	ELMo		anneaux	ceintures	sphères	balls
	BoW		anneaux	rouhault	penon	mémère
	fastText	wrestlers: 27 *lutteurs*: 10	catches	catchers	wrestlers	catch
	ELMo		lutteurs	joueurs	joueuses	joueuses
	BoW		grandidieri	bergroth	committeer	shinjitsu
Specialized	fastText	birth: 9k *naissance*: 14k	naissance	décès	âge	deuil
	ELMo		naissance	baptême	éclosion	décès
	BoW		naissance	enfant	mère	femme
	fastText	keratin: 66 *kératine*: 52	kératine	fibroblaste	adipocyte	prolactine
	ELMo		kératine	collagène	mélanine	tanin
	BoW		kératine	luminales	fibrine	hyperdensité
	fastText	vincristine: 23 *vincristine*: 15	vincristine	dominique	monique	colette
	ELMo		vincristine	raloxifène	fusarium	doxorubicine
	BoW		vinorelbine	herceptin	rechuter	docétaxel

Table 3: 4 best translations obtained for pairs on Supervised (MUSE).

6 Conclusion

This work sought to study the different BLI methods and their evaluation when using specialized and general comparable corpora. Comparing mapping methods, we observe that the results follow the same trends. The choice of seed lexicon, however, is more impactful as bigger lexicons cause the performances to decrease in the general domain and increase in the specialized one. Supervised mapping is still the way to go without needing large lexicons. FastText gives the best results on the original lists, as they are composed of a lot of graphically close words. ELMo gets better results with the sublists, as it is better at capturing concepts.

Reference lists for these approaches are often used as-is. For both domains, we challenge the validity of these lists, arguing that not all the words are worth translating. To verify this claim, we broke down our lists into sublists, isolating subsets that make more sense to translate. When comparing the results for the original list and the sublists, we see clear differences in performances, indicating the necessity of having more fine-grained reference lists.

References

1. Artetxe, M., Labaka, G., Agirre, E.: Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In: AAAI. pp. 5012–5019. New Orleans, LA, USA (2018)
2. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: ACL. pp. 789–798. Melbourne, Australia (2018)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. CoRR **abs/1607.04606** (2016)
4. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. CoRR **abs/1710.04087** (2017)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. pp. 4171–4186. Minneapolis, MN, USA (2019)
6. Hazem, A., Morin, E.: Leveraging meta-embeddings for bilingual lexicon extraction from specialized comparable corpora. In: COLING. pp. 937–949. Santa Fe, NM, USA (2018)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 (2013)
8. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR **abs/1309.4168** (2013)
9. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR **abs/1802.05365** (2018)
10. Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: ACL. pp. 519–526. College Park, MD, USA (1999)
11. Schuster, T., Ram, O., Barzilay, R., Globerson, A.: Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In: NAACL-HLT. pp. 1599–1613. Minneapolis, MN, USA (2019)
12. Zhang, Z., Yin, R., Zhu, J., Zweigenbaum, P.: Cross-lingual contextual word embeddings mapping with multi-sense words in mind. arXiv:1909.08681 (2019)