Revisiting comparable corpora in connected space

Pierre Zweigenbaum

CNRS, UPR 3251, LIMSI
91403 Orsay, France
pz@limsi.fr

Abstract

Bilingual lexicon extraction from comparable corpora is generally addressed through two monolingual distributional spaces of context vectors connected through a (partial) bilingual lexicon. We sketch here an abstract view of the task where these two spaces are embedded into one common bilingual space, and the two comparable corpora are merged into one bilingual corpus. We show how this paradigm accounts for a variety of models proposed so far, and where a set of topics addressed so far take place in this framework: degree of comparability, ambiguity in the bilingual lexicon, where parallel corpora stand with respect to this view, e.g., to replace the bilingual lexicon. A first experiment, using comparable corpora built from parallel corpora, illustrates one way to put this framework into practice. We also outline how this paradigm suggests directions for future investigations. We finally discuss the current limitations of the model and directions to solve them.

1. Introduction

The standard approach to bilingual dictionary extraction from comparable corpora (Rapp, 1995; Fung and McKeown, 1997) proposes to perform monolingual distributional analysis in each of the two comparable corpora. It represents source and target words with context vectors, and a transformation of source context words into target context words through a dictionary. Previous work has investigated variations on context vector construction (context nature and size, association scores, e.g., (Laroche and Langlais, 2010; Gamallo and Bordag, 2011)) and on the seed-dictionary-based transformation: origin and coverage of the dictionary, e.g., (Chiao and Zweigenbaum, 2003; Hazem and Morin, 2012), complementary transformations (Gaussier et al., 2004), disambiguation of dictionary entries (Morin and Prochasson, 2011; Apidianaki et al., 2013; Bouamor et al., 2013b), acquisition of the dictionary from parallel corpora (Morin and Prochasson, 2011; Apidianaki et al., 2013).

Here we want to emphasize the overall space which is created by this construction. Previous work has hinted at this overall space (e.g., (Gaussier et al., 2004)) or used it explicitly (Peirsman and Padó, 2010) but has not to our knowledge investigated further the view that it can provide on the task and its related issues. The goal of this paper is to draft a model of this space and to point at the avenues it opens for further research. Therefore this paper is a rather abstract, first stab at a description of this model, and leaves both a precise formalization and concrete experiments for further work. It also leaves for future work the handling of multi-word expressions. This type of exposition may incur risks of "hand waiving", which we have tried to minimize. Its main contributions (and outline) are the following:

- The description of a unified space embedding the context vectors of the two comparable corpora;
- The description of a connected, bilingual corpus generated from the two comparable corpora;
- A reformulation of some topics in bilingual lexicon extraction from comparable corpora;
- Suggestions for future research spawned by this unified space.

2. Related work

The introduction has shortly enumerated several dimensions of research on bilingual lexicon extraction from comparable corpora. The work closest to what we develop here is that of (Gaussier et al., 2004). A core component of the geometric view of (Gaussier et al., 2004) is the space defined by (source, target) word pairs in the bilingual dictionary. Among other things, (Gaussier et al., 2004) propose to represent words of both the source and target corpora in this common space, effectively creating a unified space. We propose below to extend this space and to study the view it gives of the joined comparable corpora.

Joint bilingual representations have been proposed in the past in various settings. Dual-language documents have been proposed by (Dumais et al., 1996), where a document and its translation are merged into a bilingual document; Latent Semantic Indexing is then performed on the collection of dual-language documents. Since we work with comparable corpora, we extend this concept to that of a dual-language corpus.

Translation pairs, i.e., bilingual dictionary entries, are used by (Jagarlamudi and Daumé III, 2010) as a substitute for 'concepts' to create cross-language topics. We also use translation pairs as basic units for cross-language representation; in our setting they are used in context vectors and in the above-mentioned dual-language corpus.

The notion of a bilingual vector space for comparable corpora, labeled with translation pairs, has already been proposed by (Peirsman and Padó, 2010). To avoid the need for a bilingual dictionary, they bootstrap translation pairs with "frequent cognates, words that are shared between two languages" (Peirsman and Padó, 2010). This creates a bilingual space in which words of each language are represented by context vectors in which context words are translation pairs. Both source and target words can be compared according to the similarity of their context vectors. Given a source word s, its nearest neighbor t in the target language is a candidate translation. (Peirsman and Padó, 2010) select

$$\begin{array}{l|c|} & \mathit{women} \\ \cdots & \vdots \\ \cdots & \vdots \\ \mathit{pregnant} & 4.394197 \\ \cdots & \vdots \\ \cdots & \vdots \end{array} \qquad E = \begin{array}{c|cccc|} & e_1 & e_j & \cdots & e_m \\ e_1 & \ddots & a(e_1, e_j) & & \\ \vdots & & \vdots & & \\ e_i & & a(e_i, e_j) & & \\ \vdots & & \vdots & & \\ e_m & & a(e_m, e_j) & & \ddots \end{array} \qquad F = \begin{array}{c|cccc|} & f_1 & f_l & \cdots & f_n \\ f_1 & \ddots & a(f_1, f_l) & & \\ \vdots & & \vdots & & \\ f_k & & a(f_k, f_l) & & \\ \vdots & & \vdots & & \\ f_n & & a(f_n, f_l) & & \ddots \end{array}$$

Figure 1: Context vectors in source and target corpora: the column for e_j (resp. f_k) represents its context vector, and $a(e_i, e_j)$ (resp. $a(f_k, f_l)$) is the association strength of e_i and e_j (resp. f_k and f_l).

candidate pairs (s, t) where t is the nearest target neighbor of s and s is the nearest source neighbor of t. Iterating this process extends the initial set of seed bilingual pairs into a larger bilingual lexicon. This notion of a bilingual vector space was only a means to an end in (Peirsman and Padó, 2010). We explore it further in the present paper.

3. Reformulating the standard approach to bilingual lexicon extraction from comparable corpora

3.1. Monolingual distributional analysis of source and target corpora

The distributional hypothesis characterizes the meaning of a word by the distribution of its usages in a language sample: a corpus. The original formulation by Harris (see details in (Habert and Zweigenbaum, 2002), citing (Harris, 1991)) relies on relations between operators and arguments. A common approximation consists in representing word usage through co-occurrence with other words in the corpus. Whatever the choice, given the vocabulary V, this associates to a given word $e_i \in V$ a vector of words $e_j \in V$ to which it is syntagmatically associated, and which is usually called its *context vector*. For example, context words (e.g., *pregnant*) in Sentence (1) contributes to the characterization of the context vector for *women* (see Figure 1, left):

(1) information for pregnant <u>women</u> and children

$$\begin{array}{l|c|} & \mathit{women} \\ \cdots & \vdots \\ \cdots & \vdots \\ [\mathit{pregnant} \sim \mathit{enceintes}] & 4.394197 \\ \cdots & \vdots \\ \cdots & \vdots \end{array}$$

Figure 2: A context vector of the source corpus, with entries translated into the target language.

Overall, this creates a word×word matrix E of dimension $|V| \times |V|$ in which $E_i^j = a(e_i, e_j)$ is the association strength of e_i and e_j. Mutual information, log-likelihood ratio, and odds-ratio, among others, are common values for this association strength (see e.g. (Evert, 2005; Laroche and Langlais, 2010) for more association scores).
Given two corpora S and T (typically, here, two comparable corpora in two different languages), composed of vocabularies V and W, we can build word×word association matrices E and F of dimensions $|V| \times |V|$ and $|W| \times |W|$ (see Figure 1, center and right).

3.2. How (unambiguous) bilingual links connect source wband target spaces

The standard approach additionally relies on a bilingual dictionary $D = \{[s_i \sim t_j]\}$, i.e., a set of [source∼target] word pairs. Its fundamental hypothesis is that word distribution reflects meaning and that meaning is preserved through translation, from which it assumes that the distribution of source words in the source corpus is similar to the distribution of their translations in the target corpus.[1] To simplify the exposition, we assume here that the dictionary introduces no ambiguity: it provides exactly one translation for the input source words that it contains (and reciprocally for target words). We do not assume that it has full coverage of the source or target corpus, otherwise there would remain no unknown word to translate.
Let us start from the context vector representation $(a(e_i, e_j))_{i=1}^{i=m}$ of a source word e_j in the source corpus, where $a(e_i, e_j)$ is the value of the vector on the axis provided by word e_i. The dictionary D is used to translate the entries in this context vector: based on translation pairs $[e_i \sim f_k] \in D$, i.e., where f_k is a translation of e_i through the dictionary, it produces a representation $(a(f_k, e_j))_{k=1}^{k=n}$ of the source word e_j in the target corpus (see Figure (2)). In this representation, the same value $a(f_k, e_j) = a(e_i, e_j)$ $= a([e_i \sim f_k], e_j)$ is assumed to represent the association that the source word e_j would have with the target word f_k translated from e_i if e_j were occurring in the target corpus. This creates a representation of the position of e_j in the target space: target words f_l whose positions are close to it are candidates to translate e_j.

[1] Note that (Harris, 1988, viii) considers that this applies to the language of a given subscience (see again (Habert and Zweigenbaum, 2002)) rather than to the whole language.

$$E_t = \begin{array}{c} \\ e_1 \\ \vdots \\ e_{m-p} \\ [e_{m-p+1} \sim f_1] \\ \vdots \\ [e_m \sim f_p] \end{array} \begin{array}{ccc} e_1 & e_j & e_m \\ \ddots & a(e_1, e_j) & \\ & \vdots & \\ & a(e_{m-p}, e_j) & \\ & a([e_{m-p+1} \sim f_1], e_j) & \\ & \vdots & \\ & a([e_m \sim f_p], e_j) & \ddots \end{array} \qquad F_t = \begin{array}{c} \\ [e_{m-p+1} \sim f_1] \\ \vdots \\ [e_m \sim f_p] \\ f_{p+1} \\ \vdots \\ f_n \end{array} \begin{array}{ccc} f_1 & f_l & f_n \\ \ddots & a([e_{m-p+1} \sim f_1], f_l) & \\ & & \\ & a([e_m \sim f_p], f_l) & \\ & a(f_{p+1}, f_l) & \\ & & \\ & a(f_n, f_l) & \ddots \end{array}$$

Figure 3: Translated context vectors in source (E_t) and target (F_t) corpora. $[e_{m-p+d} \sim f_d]_{d \in (1 \ldots p)}$ are translation pairs in the dictionary. Instead of discarding the non-translated contexts of the vectors, we keep them untouched.

$$G = \begin{array}{c|cccccccc} & e_1 & e_j & \ldots & [e_{m-p+d} \sim f_d] & f_l & \ldots & f_n \\ \hline e_1 & \vdots & a(e_1, e_j) & & \vdots & & & \\ \vdots & \vdots & \vdots & & \vdots & & \mathbf{0} & \\ e_i & \vdots & a(e_i, e_j) & & \vdots & & & \\ [e_{m-p+1} \sim f_1] & \vdots & a([e_{m-p+1} \sim f_1], e_j) & & \vdots\vdots & a([e_{m-p+1} \sim f_1], f_l) & \vdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots\vdots & \vdots & \vdots & \vdots \\ [e_m \sim f_p] & \vdots & a([e_m \sim f_p], e_j) & & \vdots\vdots & a([e_m \sim f_p], f_l) & \vdots & \vdots \\ f_{p+1} & & & & \vdots & a(f_{p+1}, f_l) & \vdots & \vdots \\ & & \mathbf{0} & & \vdots & \vdots & \vdots & \vdots \\ f_n & & & & \vdots & a(f_n, f_l) & \vdots & \vdots \end{array}$$

Figure 4: Translated context vectors G in source and target corpora, embedded in unified context space. $[e_{m-p+d} \sim f_d]_{d \in \{1 \ldots p\}}$ are translation pairs in the dictionary.

Since generally not all source and target words belong to the dictionary, only a part of a source context vector (say p entries) goes through this translation, while the rest is ignored. Let us assume for ease of exposition that we reorder the rows (and columns) of E (resp. F) with the p in-dictionary entries last (resp. first). The translated version E_t of the source (resp. F_t of the target) context vectors can then be schematized as shown in Figure 3 (we keep the out-of-dictionary part of the vectors though). This reveals the common representation subspace created by the dictionary entries ($[e_{m-p+1} \sim f_1] \ldots [e_m \sim f_p]$, in red in Figure 3).

3.3. Embedding bilingual corpora into a unified space

This common subspace provides a basis on which to merge the two sets of context vectors. Of the m dimensions of E and of the n dimensions of F, p are common to both. These vectors can thus be extended to dimension $q = m + n - p$: vectors of E_t are extended with $n - p$ zeros at their end, and vectors of F_t are extended with $m - p$ zeros at their beginning.[2] Besides, to highlight some properties of the obtained representation, we re-order the context vectors so that the columns for source and target words in the dictionary are next to each other. This is schematized on Figure 4, where the common subspace is shown in red, zero extensions are shown in blue, two in-dictionary context vectors are grouped under each $[e_{m-p+d} \sim f_d]$ header (in violet), and black shows the corpus-specific contexts. Note that only the red parts are used in the standard approach.

These in-dictionary context vectors have another interpretation at the text level. Substituting source (resp. target) words with translation pairs amounts to actually *replacing in the texts the source (resp. target) words present in the dictionary with concatenated bi-words*. For instance, depending on the dictionary, the English Sentence (1) may become as in Sentence (2 a) (the dictionary has no entry for *information* and *women*). Similarly, in the reverse direction, the French sentence *une forte proportion de femmes enceintes* may give rise to Sentence (2 b):

(2) (a) information| for|intention pregnant|enceintes women| and|et children|enfants

(b) a|une high|forte proportion|proportion of|de |femmes pregnant|enceintes

Figure 5 displays the same examples graphically, with En-

[2] Note again that we do not discard the non-translated contexts of these vectors. This contrasts to the standard approach where only the in-dictionary contexts are kept and then compared. We return to this point below.

information	for	pregnant	women	and	children
	intention	enceintes		et	enfants

a	high	proportion	of		pregnant
une	forte	proportion	de	femmes	enceintes

Figure 5: Bilingual corpus: an English sentence and a French sentence. In this example, *information*, *women*, and *femmes* are out-of-dictionary words.

glish words on top and French words at the bottom. Blue color marks the source sentence. Once transformed this way, the two comparable corpora can be merged into one bilingual corpus. To avoid confusion between source and target cognates, all out-of-dictionary words in the source and target corpora are marked with their language.[3]

The representation of words in this corpus can follow the standard distributional practice outlined in Section 3.1. Since source corpus words outside the dictionary never co-occur with target corpus words outside the dictionary, the two corresponding quadrants of the matrix in Figure 4 are filled with zeros. This should make the contribution of out-of-dictionary contexts minimal in the computation of vector similarity.

More precisely, if the dot product is used to compare context vectors, the representation in Figure 4 leads to the same results as truncating context vectors to their dictionary part, as is performed in the standard approach. However, if the similarity of two vectors is instead computed through a formula which takes into account all components of both vectors (e.g., cosine similarity normalizes the dot product by dividing it by the norms of the two vectors, and the Jaccard index divides the common features by the union of all features of the two context vectors), the formulation in Figure 4 should lead to reduced similarity values for each word with a strong association with out-of-dictionary words. If we consider that for a given word, the stronger its associations with out-of-dictionary words, the poorer the fidelity of its context vector, reducing its similarity to other context vectors might not be a bad move. This suggests a direction for new investigations.

Note also that for each $d \in \{1 \ldots p\}$, the context vectors of translation pair items e_{m-p+d} and f_d are expected to be more similar to each other than to any other context vector. These pairs of in-dictionary context vectors might thus provide a training set to tune some parameters or to train supervised methods. However, replacing e_{m-p+d} and f_d with a concatenated bi-word in the corpus replaces their two context vectors with a single one (not shown in Figure 4). This forces a single distribution on the resulting bi-word. Such merged context vectors are the only ones that may have non-zero out-of-dictionary context words in both the source and target subspaces of the corpus.[4]

To summarize, we have proposed here:

1. A unified context matrix which embeds context vectors of both source and target corpora; and
2. An associated merged bilingual corpus, some of whose "words" are bilingual word pairs.

The merged bilingual corpus has only been sketched. While computations are performed on the unified context matrix, the main intention of the merged bilingual corpus is to produce a concrete object which can support human observation and reasoning, and thereby complement the more abstract artifact of context vectors in unified context space. It is defined as a corpus whose contexts produce the unified context matrix. If the bilingual dictionary is not ambiguous (i.e., it only contains one-to-one mappings between source and target words), the merged corpus can be defined by simple substitution as in the present section. If the bilingual dictionary is ambiguous (see Section 4.3. below), creating the bilingual corpus requires a more complex management of individual contexts which goes beyond the present paper. This difficulty in building the bilingual corpus may be taken as a clue that ambiguous dictionary entries create a problem for bilingual lexicon extraction from comparable corpora, and should thus be resolved before bilingual lexicon extraction.

4. Revisiting common topics in bilingual lexicon extraction

4.1. Bilingual lexicon extraction as "a-lingual" distributional analysis and similarity

The unified context vector space contains both source and target context vectors. Similarity in this space can therefore be used to compare source and target context vectors directly, hence to look for word translations. Moreover, clustering in this space results in clusters which can contain at the same time source and target context vectors, which are similar either in source space (monolingual distributional similarity), in target space (same), or across the two (cross-lingual distributional similarity, aimed at spotting translations).

Having one unified space might be thought at first sight to help reduce the common propensity to use directional methods, which then need to be symmetrized a posteriori as in (Chiao et al., 2004). This is however not necessarily the case: even within unified space, (Peirsman and Padó, 2010) still opt to enforce symmetric conditions to select similar words.

4.2. Degree of comparability

(Déjean and Gaussier, 2002) consider that two corpora are comparable if a non-negligible subpart of the vocabulary V

[3] For instance by prefixing them with *lang_*, e.g. *en_* and *fr_*. In our experiments we adopted a simpler convention where a translation pair $[e \sim m]$ is noted $e \mid f$, and source or target out-of-dictionary words are noted respectively $e \mid$ and $\mid f$, as seen in Example (2 a) for *information* and *women*.

[4] We might also keep the original individual context vectors of e_{m-p+d} and f_d, and add to them, instead of substituting for them, their merged context vector. This amounts to duplicating the sentences (or more precisely the contexts) in which words e_{m-p+d} or f_d occur: keeping the original sentence and creating a copy where occurrences of e_{m-p+d} or f_d are replaced with $[e_{m-p+d} \sim f_d]$.

of the source corpus has a translation in the target vocabulary W and reciprocally. (Li and Gaussier, 2010) base their measure of comparability of two corpora on the proportion of words in V (resp. W) whose translations are found in W (resp. V). This proportion corresponds to the proportion of rows in the E_t or F_t matrix which could be covered by a complete dictionary—or which an oracle method could map to a correct translation in the corpus. In contrast, comparability measures which use features other than simple words translations (Su and Babych, 2012) do not have a simple counterpart in these matrices.

4.3. Ambiguity in the bilingual lexicon

The proposed construction emphasizes the importance of disambiguating dictionary word translations, which recent work (Apidianaki et al., 2013; Bouamor et al., 2013b) has shown to be able to bring substantial improvements in bilingual lexicon extraction from comparable corpora. However, if multiple translations remain for source dictionary words (e.g., $[e_{m-p+d} \sim f_{d_1}]$, $\dots [e_{m-p+d} \sim f_{d_t}]$), the context vector view presented in Section 3.3. should be adapted.

One way to handle this would be to create additional rows (and columns) in matrix G for the additional translation pairs. This amounts to duplicating the sentences (more precisely, contexts) in which the source word e_{m-p+d} occurs: each resulting sentence S_i would replace occurrences of e_{m-p+d} with $[e_{m-p+d} \sim f_{d_i}]$. However, if several source words $e_a, e_b, \dots$ map to the same target word f_d, this results in distinct representations $[e_a \sim f_d]$, $[e_b \sim f_d]$, $\dots$ of the same target word f_d which split the distribution of this target word into several parts. This could be a reasonable option if this separates distinct senses of f_d.

Another way would be to assume a less constrained mapping (typically, a linear transformation) through the dictionary from source words to target words. This can be defined by a transformation matrix M (see, e.g., (Gaussier et al., 2004)) whose row indexes are the source words that have an entry in the dictionary, whose column indexes are the target words which the dictionary proposes for at least one source word, and where $M_{ij} = 1$ (or some given positive weight, for instance such that $\sum_j M_{ij} = 1$ to encode a distribution of word translation probabilities) iff $[e_i \sim f_j]$ is in the dictionary and $M_{ij} = 0$ otherwise. As announced in Section 3.3., this method makes it more difficult to design an associated merged corpus. A direction to consider to create this merged corpus would be to include in this corpus not only full sentences, but also isolated phrases embodying elementary contexts.

All in all, the present discussion emphasizes that disambiguating source (and target) words helps obtain a better-defined model and could help design a more natural merged corpus. The methods adopted by (Apidianaki et al., 2013) look particularly relevant for this purpose since they induce clusters of translations which create sense clusters in the target corpus, hence seem compatible with the first above-mentioned way to handle ambiguity.

4.4. Parallel corpora in connected space

Parallel corpora[5] are often considered to be an ideal version of comparable corpora: they maximize comparability inasmuch as most source words can be aligned to a target word, and reciprocally. Indeed, parallel corpora also have drawbacks, the main one being that they are subject to translation bias: at least one of the two parallel corpora has been obtained by translating from a source language, and may contain calques, so the parallel corpus is a less good sample of that language. However, as in most work on parallel corpora, we shall ignore this property here.

We can represent two parallel corpora in the same way as comparable corpora in Section 3.1.: each corpus is subjected to distributional analysis to build context vectors. Then, instead of using an external bilingual dictionary, we can take advantage of word alignments to connect the two corpora. An advantage of word alignments (assuming they are correct) over using an external dictionary is that no disambiguation is necessary: each word translation is precisely valid in the context where it is found. Another advantage is that as mentioned above, most source words are aligned with some target word.

What is the use of considering parallel corpora under this view? Indeed, since most words can find translations through alignment, which is much more precise than distributional similarity, handling them as comparable corpora is not directly relevant for bilingual lexicon acquisition. However, let us examine their representation more closely.

A direct equivalent of a dictionary translation pair in parallel corpora is a pair of aligned $[e \sim f]$ words. However, a given source word may be translated as one among a set of variant words, and a set of different source words may obtain the same translation (which is useful to collect paraphrases (Barzilay and McKeown, 2001)). It may thus be beneficial to identify, among the possible translations of a given source word, those that are equivalent or closely related (Apidianaki, 2008) and those that are different (see also (Yao et al., 2012) for statistics on synonymy [equivalence] and polysemy [difference] in this context). Such sense clusters may provide a more relevant basis for translation pairs than individually aligned words in context vectors: by making (language-sensitive) word senses explicit, they should on the one hand lead to better generalization than individual words, while on the other hand differentiating different senses, thus potentially leading to better discrimination. Examining parallel corpora in the framework of unified context vector space thus naturally leads to considering questions and directions that have proved fruitful in the parallel corpus literature.

Another interest of representing parallel corpora in unified context space is that they can then be used in lieu of a dictionary to connect comparable corpora: this is the topic of the next section.

[5] In this paper we use the plural term ‘parallel corpora’ to refer to a pair of aligned corpora, to make it easier to refer to each corpus individually as the ‘source corpus’ and the ‘target corpus’. This departs from common usage where a parallel corpus (singular) refers to a corpus of bitexts.

4.5. Substituting the bilingual dictionary with a parallel corpus

Replacing the bilingual dictionary with one obtained from a pair of parallel corpora has been proposed by (Morin and Prochasson, 2011; Apidianaki et al., 2013). As explained in the previous section, parallel corpora have an advantage over a dictionary: their word alignments are found in the context of a sentence, so that the translation they show for a given (possibly ambiguous) source word in a source sentence is a correct translation of that source word in that source context, displayed in the context of the target sentence in which it occurs. In other words, parallel corpora directly implement the substitution introduced in Section 3.3. Therefore, an ideal situation when using parallel corpora would be to add them to the comparable corpora, thereby directly connecting the source and target corpora. For consistency, the parallel corpora should be in-domain, i.e., the source (resp. target) parallel corpus should be comparable to the source (resp. target) comparable corpus.

However, (Morin and Prochasson, 2011) and (Apidianaki et al., 2013) kept their parallel corpora separate from the comparable corpora. (Morin and Prochasson, 2011) used in-domain parallel corpora but discarded them after obtaining a dictionary of aligned words. (Apidianaki et al., 2013) used out-domain parallel corpora, induced word senses from them, and used these sense clusters plus information from the parallel corpora to disambiguate translations. This makes better use of the observed word distributions in the parallel corpora. Still, a step further in this direction would consist in extending the latter method by using in-domain parallel corpora: applying (Apidianaki et al., 2013)'s method to induce word senses and to translate context vectors, passing to unified context space, and adding the parallel corpora to unified context space as explained in Section 4.4.

When in-domain parallel corpora are scarce, they can be generated by machine translation from a part of the comparable corpus (Abdul-Rauf and Schwenk, 2009). Assuming that the machine translation system used to do so has been trained on a large pair of parallel corpora for the considered language pair, this creates a chain of steps which propagate translation pairs: (i) translation pairs are learned from large (out-domain) parallel corpora into the phrase table; (ii) they are used to produce (artificial) (in-domain) parallel corpora by translating existing sentences of the comparable corpora (note that this can be done in both directions); (iii) translation pairs instantiated in the artificial parallel corpora link the two comparable corpora; (iv) distributional analysis and similarity in the comparable corpora suggest new translation pairs. Some amount of loss is to be expected at each stage: as in many other directions listed in this paper, experiments will be useful to know to which extent this impedes the outlined method.

5. A preliminary experiment

As a preliminary, controlled experiment, we performed translation spotting in unified space in a pair of comparable corpora. We created these comparable corpora in such a way that many of their words come with tailored, low-ambiguity translations. We started from English-French parallel corpora obtained from the *Health Canada* bilingual Web site (Deléger et al., 2009) and re-used by (Ben Abacha et al., 2013) for cross-language entity detection. The corpus was word-aligned with Fast Align (Dyer et al., 2013) in forward and reverse directions, then symmetrized with atools with the grow-diag-final option. It was then split into two halves in the order of the files (hence the topics covered by the two halves are expected to show some differences). The first half was used as an English source corpus (with French translation), and the second half as a French source corpus (with English translation).

When a source word was aligned to multiple target words, a more selective word alignment was obtained by computing an association score (discounted log odds ratio) over the word alignment links and keeping the link with the most associated target word. Links under a threshold were also discarded (we selected a threshold of 1 based on initial experiments). The target word selected this way was considered to be the translation of the source word and was pasted to it to create a bi-word as per the notations showed in Sentences 2 a and 2 b in Section 3.3. (see also Figure 5). This created two artificial comparable corpora. In each of these two corpora, some source words were mapped to target words as though through a dictionary—actually thanks to the word alignment process.

We then simulated out-of-dictionary words by surgically removing some of these translations. Given a translation pair $[e \sim f]$, in the English corpus we modified all bi-words $e|*$ into $e|$ and all bi-words $*|f$ info $|f$; in the French corpus we did the same in the opposite order. The examples cited in Section 3.3. were actually extracted from this corpus; they were obtained by removing the translation pairs [*women* ∼ *femmes*] and [*information* ∼ *information*] from the two parts of the corpus. We did this for several series of translation pairs: 31 among the most frequent ones, 54 at rank 1000, 45 at rank 5000, 48 at rank 10000, and 49 at rank 15000, for a total of 227 translation pairs. After this operation, the two halves of the corpus were pasted together, thus producing one bilingual corpus with 2×227 additional out-of-dictionary words (slightly less actually since our sample of translation pairs happened to include a few common source or target words). This corpus contains 2.1 million words.

We then performed distributional analysis of this corpus in unified space: we built context vectors for each (bi)word in the corpus (minimum 5 occurrences, stop-word removal in both languages, window of 5 words left and right, discounted log odds-ratio as in (Laroche and Langlais, 2010)). Context vectors were truncated to the 1000 most associated context words. Vector similarity was computed by taking the cosine of the two vectors (we also tested the dot product).

We performed the translation spotting task by taking as source words the above 227 pairs of artificial out-of-dictionary words. For each source word, we retrieved the corresponding context vector, computed its similarity to all other context vectors, and ranked them in descending similarity order (we kept up to 500 most similar context vectors). We evaluated the results by checking whether the word with the closest context vector was the refer-

	sim	dir	F-measure
success@1	cos	f→e	0.3982
		e→f	0.4398
	dot	f→e	0.5113
		e→f	0.4213
success@o1	cos	f→e	0.6833
		e→f	0.7083
	dot	f→e	0.6606
		e→f	0.6806

Table 1: Translation spotting in unified space. N=227 test pairs in either direction; sim = similarity: cos = cosine, dot = dot product; dir = direction of translation.

ence translation (the other word of the translation pair), e.g. whether starting from $women|$, the closest context vector was that for $|femme$ (*success@1*). Sometimes the closest context vector may represent a word of the same language. Therefore we also performed the same check restricted to out-of-dictionary words of the other language (*success@o1*, where *o* stands for out-of-dictionary and also for other). This second measure can be seen as more realistic since we have this knowledge and can use it anyway in a translation spotting task. However, out-of-dictionary words include on the one hand natural OOD words which could not be aligned reliably when preparing the corpus, and on the other hand artificial OOD words which can have a different distribution. This may bias their recognition and lead to an optimistic evaluation. Hence our trying to reduce this bias by selecting words in a variety of frequency ranges.

Table 1 displays the obtained results. A detailed analysis of this first experiment is beyond the scope of this paper; we may observe nevertheless that success@1, between 0.40 and 0.51, would be rather high for comparable corpora, and that success@o1, between 0.66 and 0.71, is as expected much higher but probably optimistic. The important point is that this exemplifies distributional analysis in unified space, where the translation links which create biwords are obtained from parallel corpora instead of a preexisting dictionary. The extension of this experiment by adding non-parallel texts to a parallel kernel is left for future work.

6. Embedding space suggests directions for future investigations

Presenting the unified context space and the connected bilingual corpus led us to mention several topics about bilingual lexicon acquisition from comparable corpora which deserve investigation. Among others we mentioned keeping whole context vectors in similarity computation instead of truncating their out-of-dictionary part; performing similarity computation directly on unified context space; performing cross-language clustering on unified context space; whether or not to merge the context vectors of in-dictionary words, and its consequence on bilingual lexicon extraction; connecting parallel corpora to unified context space; exploring the relevance of creating them through machine translation.

The handling of the context vectors of in-dictionary words, with a source view (see the violet e_{m-p+d} column in Figure 4), a target view (violet f_d column), and possibly a merged view (not shown on the figure), is reminiscent of the feature augmentation proposed by (Daumé III, 2007) to help domain adaptation. The parallel here would be that the merged context vectors of in-dictionary words could help connect word distributions in the two “domains” (here languages), for instance when computing cross-language word clusters on unified context space.

As an application, bilingual word classes obtained through cross-language clustering can provide additional data for methods such as (Täckström et al., 2012) which aim at direct transfer of NLP components from one language to another.

How to create a merged bilingual corpus when multiple translations are provided for some words in the dictionary has been left undetermined in the above sections. A word lattice representation (more exactly, a directed acyclic graph) encoding alternative words could help solve the problem. The translation pair representation adopted in this paper would then be extended to pairs of disjunctions of words. However, this is likely to amount to merging the target (resp. source) word distributions for all alternate translations, which should be separated at least into sense clusters (see Sections 4.4. and 4.5. above).

7. Relation to non-standard methods of bilingual lexicon extraction from comparable corpora

The present work focuses on the above-mentioned ‘standard approach’ to bilingual lexicon extraction from comparable corpora. (Déjean et al., 2002) have proposed to extend this method by representing words through their distributional similarity to the terms of a bilingual thesaurus. That is, instead of using context vectors to represent words directly, they use context vectors to compare words to the entries of a bilingual dictionary (more precisely a thesaurus of the domain), itself represented by the context vectors of its terms as computed in the corpus. Words are thus represented by vectors of similarity values to the dictionary. The source and target parts of their comparable corpora are still used to compute context vectors, but in this method they are used as intermediate representations to obtain the similarity vectors. Since this extended method also relies on a bilingual dictionary used to translate terms occurring in the corpus, it is also a possible candidate to submit to the reformulation that we propose below. However, its bilingual dictionary is actually a thesaurus where multiword terms are a majority, and (Déjean et al., 2002)’s method does not require these multiword terms to occur as a unit: this is an obstacle to the reformulation we proposed for the standard method.

Instead of using distributional similarity in local contexts and a bilingual dictionary, some bilingual lexicon extraction methods use bilingual pairs of documents. This is the case of (Bouamor et al., 2013a) who, following (Gabrilovich and Markovitch, 2007)’s Explicit Semantic Analysis (ESA) method, represent a word by the vector of

Wikipedia pages in which it occurs. Inter-language links identify pairs of pages which describe the same entry in different languages. (Bouamor et al., 2013a) follow these links to 'translate' source ESA vectors into target ESA vectors, and then to identify candidate translations of the source word. Wikipedia is arguably a comparable corpus, but knowledge of the comparability (and often the translation) of document pairs is used here as a replacement for the bilingual dictionary; the method does not rely on an external pair of comparable corpora. And since translation takes place at the level of whole documents (the Wikipedia pages) rather than at the level of individual words in the texts, it seems difficult to submit it to our reformulation.

Beyond bilingual extraction from comparable corpora, a reference set of parallel documents (called "anchor texts") is also used by (Forsyth and Sharoff, 2014): it serves as a base to compute the vector of similarities (a similarity profile) of a text to every document in the set. Having translations of each base document enables the authors to use the same device as in bilingual lexicon extraction through a bilingual dictionary: the similarity profile of a text in a source language can be 'translated' to a target language and compared to similarity profiles of texts in the target language, hence computing inter-text similarities across languages. Again we find here the principle of multilingual linkage at the level of whole documents.

8. Current limitations and future work

As announced in the introduction, this paper is a first sketch of a renewed framework for studying bilingual lexicon extraction from comparable corpora. It takes a simple form when a one-to-one dictionary is used, which is the case in a large subset of the comparable corpora literature, where often the first or most frequent translation is used alone. However, when multiple translations are taken into account, we have seen that details of the representation need to be worked out.

The main limitation of the present paper is its double lack of a precise formalization and of experiments, which are left for further work. We believe it may be productive however to give early exposure of the above principles to public scrutiny, rather than deliver them piecewise with accompanying formalization and experiments. The first experiment presented in this paper, using comparable corpora built from parallel corpora, illustrates one way to put this framework into practice.

We plan to continue oracle experiments with controlled corpora, to better study the properties of the unified context space and of the merged bilingual corpus. For instance, even more constrained than the experiment of Section 5. with parallel corpora, two pseudo-comparable corpora can be built by splitting a monolingual corpus into two halves and tagging each token in each half to mark its language (say $source|$ and $|target$ as in Section 5.). This creates two comparable corpora in two 'distinct' languages. Then a varying proportion of the words w_d can play the role of in-dictionary words by entering the pairs $[source| \sim |target]$ into the dictionary, while the rest of the words are kept distinct.[6] The ability to spot pseudo-translations in various settings can then be evaluated, without interfering with issues linked to multiple dictionary translations.

9. References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. Exploiting comparable corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora*, BUCC '09, pages 46–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marianna Apidianaki, Nikola Ljubešić, and Darja Fišer. 2013. Vector disambiguation for translation extraction from comparable corpora. *Informatica (Slovenia)*, 37(2):193–201.

Marianna Apidianaki. 2008. Translation-oriented word sense induction based on parallel corpora. In Nicoletta Calzolari, Bente Maegaard Khalid Choukri, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Asma Ben Abacha, Pierre Zweigenbaum, and Aurélien Max. 2013. Automatic information extraction in the medical domain by cross-lingual projection. In *Proceedings IEEE International Conference on Healthcare Informatics 2013 (ICHI 2013)*, Philadelphia, USA, September. IEEE.

Dhouha Bouamor, Adrian Popescu, Nasredine Semmar, and Pierre Zweigenbaum. 2013a. Building specialized bilingual lexicons using large scale background knowledge. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 479–489, Seattle, Washington, USA, October. Association for Computational Linguistics.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013b. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 759–764, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The effect of a general lexicon in corpus-based identification of French-English medical word translations. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *Proceedings Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 397–402, Amsterdam. IOS Press.

[6] This creation of pseudo-translations is the reverse of the pseudo-words used in word sense disambiguation (Gale et al., 1992), which concatenate two existing words in the same language then expect a system to separate the distributions of the two original words.

Yun-Chuang Chiao, Jean-David Sta, and Pierre Zweigenbaum. 2004. A novel approach to improve word translations extraction from non-parallel, comparable corpora. In *Proceedings International Joint Conference on Natural Language Processing*, Hainan, China. AFNLP.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. 2009. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701. Epub 2009 Mar 9.

Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Working Notes of the Workshop on Cross-Linguistic Information Retrieval, SIGIR*, pages 16–23, Zurich, Switzerland. ACM.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*. Numéro spécial Alignement lexical dans les corpus multilingues, resp. Jean Véronis.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th COLING*, Taipei, Taiwan, 24 August–1 September.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart.

Richard S. Forsyth and Serge Sharoff. 2014. Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing*, 29(1):6–22.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from parallel corpora. In *Proceedings Fifth Annual Workshop on Very Large Corpora*, pages 192–202. ACL.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54–60.

Pablo Gamallo and Stefan Bordag. 2011. Is singular value decomposition useful for word similarity extraction? *Language Resources and Evaluation*, 45(2):95–119.

Éric Gaussier, J.M. Renders, I. Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 526–533, Barcelona, Spain, July.

Benoît Habert and Pierre Zweigenbaum. 2002. Contextual acquisition of information categories: what has been done and what can be done automatically? In Bruce E. Nevin and Stephen M. Johnson, editors, *The Legacy of Zellig Harris: Language and information into the 21st Century – Vol. 2. Mathematics and computability of language*, pages 203–231. John Benjamins, Amsterdam.

Zellig Sabbetai Harris. 1988. *Language and information*. Columbia University Press, New York.

Zellig Sabbettai Harris. 1991. *A theory of language and information. A mathematical approach*. Oxford University Press, Oxford.

Amir Hazem and Emmanuel Morin. 2012. Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *LREC 2012, Eigth International Conference on Language Resources and Evaluation*, pages 288–292, Istanbul, Turkey. ELRA.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, ECIR'2010, pages 444–456, Berlin, Heidelberg. Springer-Verlag.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 617–625, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Li and Éric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 644–652, Beijing, China, August. Coling 2010 Organizing Committee.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34, Portland, Oregon, June. Association for Computational Linguistics.

Yves Peirsman and Sebastian Padó. 2010. Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929, Los Angeles, California, June. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translation in non-parallel texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, student session*, volume 1, pages 321–322, Boston, Mass.

Fangzhong Su and Bogdan Babych. 2012. Development and application of a cross-language document comparability metric. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 477–487, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xuchen Yao, Benjamin Van Durme, and Chris Callison-Burch. 2012. Expectations of word sense in parallel corpora. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 621–625, Stroudsburg, PA, USA. Association for Computational Linguistics.