# Cross-Lingual Contextual Word Embeddings Mapping with Multi-Sense Words in Mind

Zheng Zhang, Ruiqing Yin, Jun Zhu, Pierre Zweigenbaum

# CROSS-LINGUAL CONTEXTUAL WORD EMBEDDINGS MAPPING WITH MULTI-SENSE WORDS IN MIND

A PREPRINT

**Zheng Zhang***
LIMSI, CNRS,
LRI, Univ. Paris-Sud, CNRS,
Université Paris-Saclay
Orsay, France
zheng.zhang@limsi.fr

**Ruiqing Yin***
LIMSI, CNRS,
Université Paris-Saclay
Orsay, France
ruiqing.yin@limsi.fr

**Jun Zhu***
CentraleSupélec
Université Paris-Saclay
Gif-sur-Yvette, France
jun.zhu@centralesupelec.fr

**Pierre Zweigenbaum**
LIMSI, CNRS,
Université Paris-Saclay
Orsay, France
pz@limsi.fr

September 20, 2019

**ABSTRACT**

Recent work in cross-lingual contextual word embedding learning cannot handle multi-sense words well. In this work, we explore the characteristics of contextual word embeddings and show the link between contextual word embeddings and word senses. We propose two improving solutions by considering contextual multi-sense word embeddings as noise (removal) and by generating cluster level average anchor embeddings for contextual multi-sense word embeddings (replacement). Experiments show that our solutions can improve the supervised contextual word embeddings alignment for multi-sense words in a microscopic perspective without hurting the macroscopic performance on the bilingual lexicon induction task. For unsupervised alignment, our methods significantly improve the performance on the bilingual lexicon induction task for more than 10 points.

***K*eywords** Contextual word embeddings · Cross-lingual mapping · ELMo

## 1 Introduction

Cross-lingual word embeddings (CLWEs), vector representations of words in multiple languages, are crucial to Natural Language Processing (NLP) tasks that are applied in multilingual scenarios, such as document classification, dependency parsing, POS tagging, named entity recognition, super-sense tagging, semantic parsing, discourse parsing, dialog state tracking, entity linking, sentiment analysis and machine translation (Ruder u. a., 2017).

Cross-lingual word embedding learning models can be categorized into three groups based on when alignment data is used: corpus preparation, training and post-training. For post-training models, research about the mapping of state-of-the-art pre-trained monolingual word embeddings across different languages (Mikolov u. a., 2013a; Joulin u. a., 2017; Peters u. a., 2018; Devlin u. a., 2019) keeps evolving with the progress of monolingual word embedding learning (Mikolov u. a., 2013b; Conneau u. a., 2017; Lefever und Hoste, 2009; Schuster u. a., 2019).

With the most recent progress of word embeddings learning by using pre-trained language representation models such as ELMo (Peters u. a., 2018), BERT (Devlin u. a., 2019) and XLNet (Yang u. a., 2019). Word embeddings move from

*Equal contribution

context-independent to contextual representations. Peters u. a. (2018) have shown that contextual word embeddings have a richer semantic and syntactic representation. For consistency and simplicity, we define two kinds of representations as word type embedding and token embedding.
**Word type embedding** Context-independent embedding of each word. Only one embedding is created for each distinct word in the training corpus.
**Token embedding** Contextual word embedding of each token. A token is one of the occurrences of a word (type) in a text, its embedding depends on its context. As a result, a word in the training corpus receives as many embeddings as its occurrences in that corpus.

Despite many advantages of token embeddings, mapping independently pre-trained token embeddings across languages is challenging: most existing word embeddings and cross-lingual mapping algorithms are based upon word type embeddings. How to apply previous cross-lingual word embedding mapping algorithms to multi-sense word embeddings remains unclear.

Schuster u. a. (2019) proposed the current state-of-the-art solution to this problem by conflating the multiple token embeddings of one word type into one context-independent embedding *anchor*, which enables word-type-based cross-lingual word embedding learning algorithms to apply to token embeddings. In their paper, the conflation of token embeddings is simply obtained by averaging them.

Although experiments show that this simple average anchor calculation is effective for cross-lingual token embeddings mapping, i.e. it obtained a better score on dependency parsing tasks than the previous state-of-the-art method, we believe there is still room for improvement, especially for multi-sense words.

Schuster u. a. (2019) found that token embeddings for each word are well separated like clouds, and the token embeddings of a multi-sense word may also be separated according to different word senses inside each token embedding cloud.

Based on these findings, we argue that averaging is not an optimal choice for multi-sense word anchor calculation, which directly influences cross-lingual token embeddings learning.

- For the supervised mapping methods (Mikolov u. a., 2013b; Xing u. a., 2015), the average anchor of a multi-sense word depends on the frequency of the token embeddings of each word sense. Besides, as each translation pair containing multi-sense words in the supervision dictionary may only cover one sense at one time, using only one anchor for each multi-sense word may not correspond to mono-sense based translation pairs.
- For the unsupervised cross-lingual word embedding learning model MUSE (Conneau u. a., 2017), because a multi-sense word may not have a translation word that would exactly have all its senses, the average anchor of that word may not find a corresponding average anchor embedding in the target language.

**Our contributions** The main contributions of this paper are the following:

- Analyze the geometric distribution of token embeddings of multi-sense words, suggesting its relation to sense embeddings.
- Using average anchor embeddings for both supervised and unsupervised cross-lingual word embedding learning models to show the existing problem.
- Propose our solutions of treating multi-sense word anchor embeddings as noise and replacing word anchor embeddings with cluster-level average anchor embeddings.

## 2 Related Work

The learning method of (Aldarmaki und Diab, 2019) relies on using parallel sentences either to generate a dynamic dictionary of token embeddings as the word-level alignment data or to calculate sentence embeddings as the sentence-level alignment data. Schuster u. a. (2019) proposed to conflate the token embeddings for each word into one anchor embedding so as to apply previous cross-lingual word embedding learning algorithms, In the following, we focus on the solution of Schuster u. a. (2019) as it does not need additional alignment data and it aims to connect all previous cross-lingual word embedding learning algorithms to the token embeddings field.

Below we introduce two cross-lingual word embedding learning methods along with their adaptations for token embeddings proposed by Schuster u. a. (2019).

### 2.1 Supervised Mapping

Supervised mapping methods aim to learn a linear mapping using the supervision of alignment data. Mikolov u. a. (2013b) introduced a model that learns a linear transformation between word embeddings of different languages by minimizing the sum of squared Euclidean distances for the dictionary entries. Based on this work, Xing u. a. (2015) proposed an orthogonal transform to map the normalized word vectors in one or both languages under the constraint of the transformation being orthogonal because of two inconsistences in (Mikolov u. a., 2013b):

- During the skip-gram model training stage, the distance measurement is the inner product of word vectors according to the objective function while the cosine similarity is usually used for word embedding similarity calculation (e.g. for the WordSim-353 task).
- The objective function of the linear transformation learning step (Mikolov u. a., 2013b) uses the Euler distance. But after mapping, the closeness of bilingual words is measured by the cosine similarity.

Xing u. a. (2015)'s experiments showed that normalized word vectors have a better performance in the monolingual word similarity task WordSim-353 and that the proposed method performs significantly better in the word translation induction task than (Mikolov u. a., 2013b).

**Adaptation for token embeddings** Given a dictionary used for supervised cross-lingual context-independent word (word type) embedding learning, Schuster u. a. (2019) proposed to generate average token embeddings anchors and to assign word anchor vectors to dictionay words.

$$\overline{e}_i = \mathbb{E}_c \left[ \boldsymbol{e}_{i,c} \right] \tag{1}$$

As shown in Equation 1, the anchor embedding of word $i$ is defined as the average of token embeddings over a subset of the available unlabeled data, where $e_{i,c}$ is the token embedding of word $i$ in the context $c$.

### 2.2 Unsupervised Mapping: MUSE

MUSE (Multilingual Unsupervised and Supervised Embeddings) is a Generative Adversarial Net (GAN)-based method and open-source tool introduced by Conneau u. a. (2017). In their paper, a discriminator is trained to determine whether two word embeddings uniformly sampled from the $50,000$ most frequent words either come from the $WS$ (aligned source word embeddings, where $S$ is the source word embeddings and $W$ is the linear transformation matrix) or $T$ (target word embeddings) distributions. In the meantime, $W$ is trained to prevent the discriminator from doing so by making elements from these two different sources as similar as possible. Besides, they defined a similarity measure, Cross-domain Similarity Local Scaling (CSLS), that addresses the hubness problem (i.e., some points tend to be nearest neighbors of many points in high-dimensional spaces), and serves as the validation criterion for early stopping and hyper-parameter tuning.

**Adaptation for token embeddings** Schuster u. a. (2019) also proposed another adaptation on top of the MUSE model (Conneau u. a., 2017) by using anchor embeddings: as they did in the supervised case, anchor embeddings are assigned as the vector representations for words. Then they use them in the unsupervised MUSE model.

## 3 Average Anchor Embedding for Multi-sense words

Using the average for anchor calculation is based on two findings from Schuster u. a. (2019)'s exploration of token embeddings:

1. The clouds of token embeddings of each word are well separated.
2. (a) The clouds of multi-sense words may be separated according to distinct senses.
   (b) Although the distances between token embeddings and the averaged token embedding cloud center are slightly larger than in single-sense words, the token embeddings of multi-sense words *still remain relatively close to their [...] anchor*. Because of this, the authors believe "these anchors can still serve as a good approximation for learning alignments".

In our opinion however, there is no reason for the distance between token embeddings of distinct senses to be small. Take the English word *bank* as an example, which has multiple distinct senses including the meaning of a financial

institution and the meaning of the river side. There is no reason why token embeddings related to the financial institution meaning should be close to token embeddings of the river side meaning.

We decided to investigate these claims by analyzing monolingual and aligned cross-lingual token embeddings. Our empirical investigation is consistent with the first conclusion (1) and the first point of the second conclusion (2a), but disagrees with the second point of the second conclusion (2b). Additionally, we attempt to explain why this second point is not likely to hold in principle.

### 3.1 Token Embeddings

To show the difference of token embedding geometrical distributions between multi-sense words and single-sense words, we need a multi-sense word that is directly related to single-sense words. The English word *lie* could be a good choice: the verb *lie* has two distinct senses, and each sense has a different past tense: *lied* (did not tell the truth) or *lay* (was in a horizontal position, was located). Besides, the English word *lie* can also be a noun, whose antonym is *truth*.

So we visualize the embeddings of the English word *lie* along with its two past tenses *lied* and *lay* and one of its antonyms, *truth*.

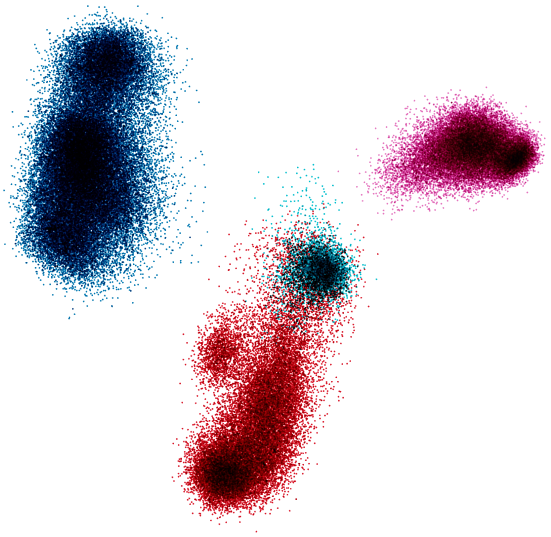

Figure 1: Token embeddings of the English word *lie* (red points, bottom middle) along with its two past tenses *lied* (light blue points, top middle) and *lay* (dark blue points, top left) and one of its antonyms *truth* (purple points, top right).

As shown in Figure 1, we found that the point clouds of the single-sense words *lied* and *truth* are more concentrated than for the multi-sense word *lie*. The point cloud of the word *lie* can be visually categorized into 3 clusters: one that overlaps the cloud of *lied* in light blue, one at the bottom, and another one on the left. By randomly selecting points and checking their corresponding sentences (Table 1) from each cluster, we found that the point clouds of the word *lie* are separated according to its distinct senses. Surprisingly, we also found that the point cloud of the word *lay* is also visually separated into 2 parts. By checking the corresponding sentences, We found the bottom part is used as the past tense of the word *lie*s and the top part is used as an adjective. (Three corresponding sentences: *In 1980, Mr. Greg Coffey was appointed the first lay principal of the College.*, *Conwell took up the post at an advanced age, and spent much of his time there feuding with the lay trustees of his parishes, especially those of St. Mary's Church in Philadelphia.* and *This includes a wide range of relationships, from monastic (monks and nuns), to mendicant (friars and sisters), apostolic, secular and lay institutes.*).

Similar findings can be found in the token embeddings of other words, in different languages and also in aligned cross-lingual embedding spaces. As suggested by Schuster u. a. (2019)'s conclusion, point clouds for each word are well separated (Conclusion 1). Besides, the point clouds of multi-sense words are also separated according to distinct senses (Conclusion 2a).

| Cluster position | Sentence | Semantic category |
|---|---|---|
| overlapping | *Yutaka and Lady Elizabeth come to the hearing and **lie** to incriminate Oyuki.*<br>*As a result of his confession, prosecutors decided not to pursue a prosecution against the remaining 20 charges, and asked that they **lie** on file, in order to spare a jury the horror of having to watch graphic images and videos of child abuse since the 71 charges which Huckle admitted to would be sufficient for a lengthy sentence.* | [verb] to deliberately say sth that is not true |
| bottom | *The city's prime locations **lie** within a radius of 6 km from Thammanam, making it thus a predominantly residential and small commercial area with basic facilities in and around the region.*<br>*As of 2009, the most heavily trafficked segments of NY 31 **lie** in and around the city of Rochester.* | [verb] to be in a particular position |
| left | *James Murphy later admitted that this was entirely a **lie** on his part, and that he does not actually jog.*<br>*The dater then asks the suitors questions which they must answer while hooked up to a **lie** detector, nicknamed the "Trustbuster".* | [noun] sth you say that you know is not true |

Table 1: Corresponding sentences selected from each visual clusters of the token embeddings of the word *lie*

### 3.2 Average Anchor Embeddings for Multi-sense Words

To analyze multi-sense word token embeddings and their average anchors in detail, we manually selected 4 multi-sense English words from the Wikipedia list of true homonyms from different perspectives:

- Distinct senses of the same part of speech (POS) (noun): **bank**-financial, **bank**-river, etc.; **spring**-season, **spring**-fountain, **spring**-coiled, etc.
- Distinct senses of different POS: **check**/Noun **check**/Verb; **clear**/Adj, **clear**/Verb

**Distribution of token embeddings for multi-sense words.** We firstly calculate all the token embeddings of the selected words over the whole English Wikipedia. We use the output of either the first or second LSTM layer of ELMo as input to the visualization (see Figure 2).

**Position of anchor embeddings for multi-sense words.** Besides the embeddings projection, we also calculate anchor embeddings for the selected multi-sense words. Then we label the 100 nearest neighbors of each anchor in the token embedding space (see the right side of Figure 2). Note that all token embeddings are also present in that visualization, but only the top 100 are labeled with the word.

**Context of token embeddings.** Also, to verify that token embeddings are geometrically separated according to distinct senses, for each cluster in the point cloud of a multi-sense word, we randomly select two points (token embeddings) in this cluster and show their corresponding sentences (see Appendix). Note that we do not apply any clustering algorithm here, clusters are just recognized based on human judgment.

**Observation.** As shown in the right side of Figure 2, most of the 100 token embeddings nearest to the anchor embedding are located in only one of the word sense clusters. The anchor is pulled closer to the sense clusters that have more token embeddings because of the averaging, which causes the first problem for cross-lingual token embeddings mapping:
**Problem 1** The anchor of a multi-sense word is biased by the frequency of the token embeddings of its senses.

### 3.3 Muti-sense Words in Dictionaries for Supervised Mapping

The supervised model is trained on a bilingual dictionary of source-target words. Dictionaries are not always generated with attention paid to multi-sense words. When a dictionary contains incomplete translation pairs related to a multi-sense word, it may contribute inaccurate mapping supervision data.

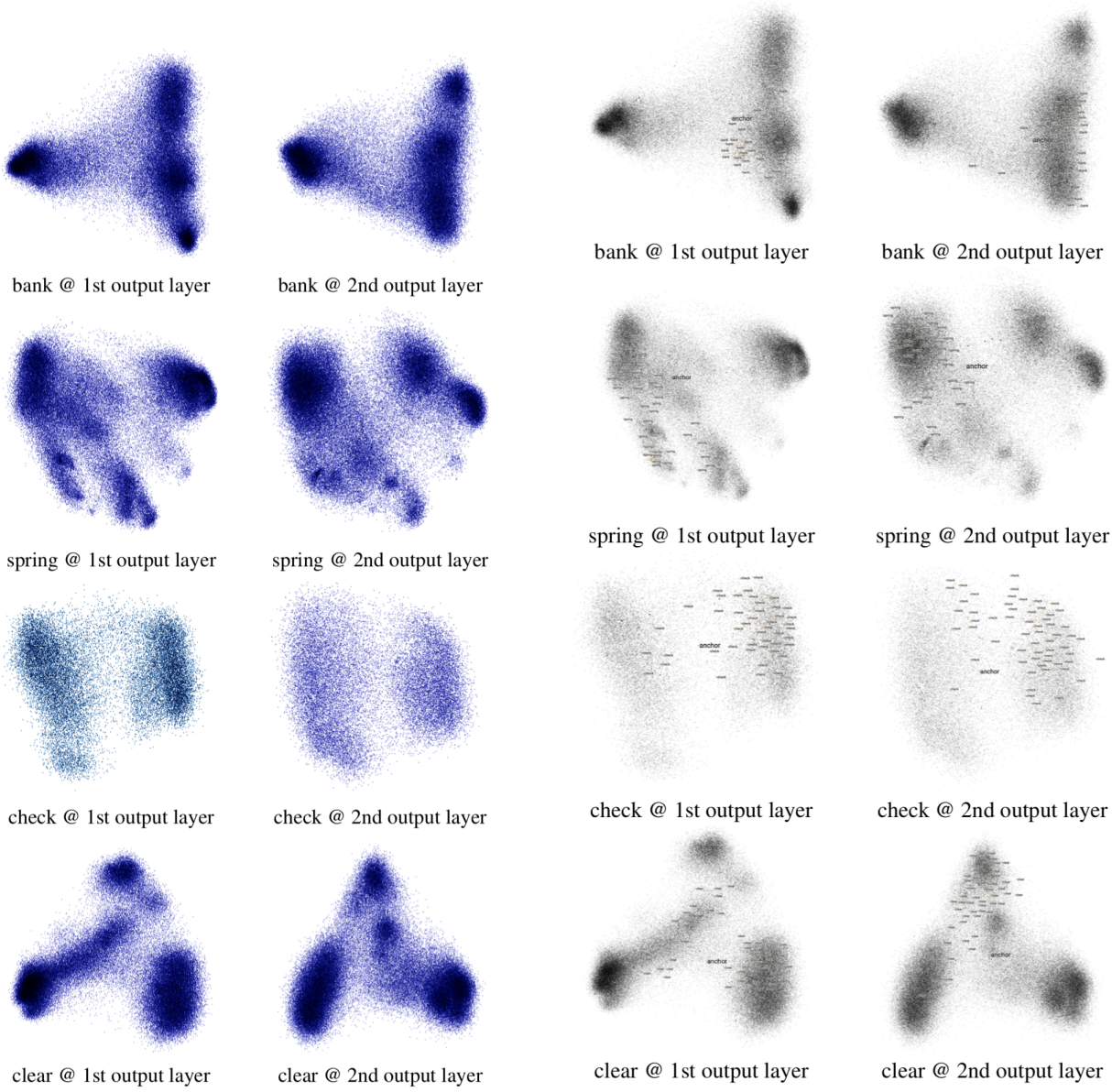


Figure 2: Token embeddings of English words *bank*, *spring*, *check* and *clear* generated from the first and second LSTM layers of the ELMo model. Labelling of the anchor embeddings (*anchor*) of English words *bank*, *spring*, *check* and *clear* and of their 100 nearest token embeddings (*bank*, *spring*, etc.). Embeddings are generated from the first and the second LSTM layers of the ELMo model.

Let us take as an example the English-French dictionary, containing 5,000 source words, used for the supervised baseline model in MUSE. We list in Table 2 all translation pairs in that dictionary related to a common multi-sense word: *bank*.

| | |
|---|---|
| bank | banques |
| bank | banque |
| banks | banques |
| banking | banques |
| banking | banque |
| banking | bancaire |

Table 2: All translation pairs related to the multi-sense word *bank* in the English-French dictionary used in MUSE for supervised mapping.

It is obvious that all translation pairs listed above are related to the *financial institution* meaning of the word *bank*. The other senses of bank, such as *land at river's side*, are ignored. Similar cases can be found for other multi-sense words in the dictionary.

**Problem 2** Because the average anchor for a multi-sense word can be considered as a general representation of all its distinct senses, using this for semantically incomplete translation pairs in a dictionary may lead to inaccurate mappings.

### 3.4 Muti-sense Words for the Unsupervised Mapping in MUSE

The unsupervised mapping model in MUSE uses a GAN to learn a linear mapping between source and target embeddings without parallel supervision data. Based on the intuition that source and target embedding spaces should share a similar global geometric structure, in the best case, source words should be mapped to their corresponding translation words in target languages.

**Problem 3** For multi-sense words, translations that have exactly the same set of senses may not exist, e.g. for the English word *bank*, there is no corresponding French word which has both the *financial institution* ("banque") and *land at river's side* ("berge", "bord", "rive", etc) senses. Therefore a multi-sense word anchor may not have a corresponding point in the target language.

## 4 Cross-lingual Token Embeddings Mapping with Multi-sense Words in Mind

We propose below solutions to these problems for both supervised mapping and unsupervised mapping methods.

### 4.1 Noise in Dictionary for Supervised Mapping

We consider incomplete translation pairs of multi-sense words as noise in the supervision data (dictionary). A simple but effective solution is to remove noise. Here we propose two types of removal:

**Form-based removal:** remove translation pairs that contain the exact multi-sense words. For instance, given that the source word *bank* is known to have multiple senses, *bank banques* and *bank banque* should be removed in Table 2.

**Lemma-based removal:** remove translation pairs containing words having the same lemma as multi-sense words. In the *bank* example, all 6 translation pairs in Table 2 should be removed as *bank*, *banks*, and *banking* have the same lemma.

Note that we do not supply a part of speech (POS) tag to the lemmatizer as there is no context to analyze the POS for words in the translation pairs of the dictionary.

### 4.2 Noisy Points for Unsupervised Mapping in MUSE

As discussed before, the exact corresponding senses-to-senses translation of a multi-sense word may not exist in target languages, i.e. the average anchor for multi-sense words may not be correctly aligned to target embedding spaces.

In that context, we consider multi-sense word anchors as noise for the unsupervised mapping model in MUSE. So we remove all multi-sense word anchors from the independently pre-trained monolingual word embeddings used for training (We name this method **anchors removal** in Table 3).

### 4.3 Cluster-level Average Anchor Embeddings for Unsupervised Mapping in MUSE

We apply the spectral clustering algorithm Wang u. a. (2018) to token embeddings of multi-sense words and calculate an average anchor embedding for each cluster. Then for each multi-sense word, we replace its average anchor embedding with cluster-level average anchor embeddings. (We name this method **anchors replacement** in Table 3.)

## 5 Experiments

### 5.1 Token Embeddings

**Pre-trained model** We use the same ELMo models as in (Schuster u. a., 2019), which are trained on Wikipedia dumps with the default parameters of ELMo (Peters u. a., 2018).
**Corpus** The Wikipedia dumps we used for specific words analysis are the same as the training data for ELMo models.
**Lexicon induction evaluation** Following (Schuster u. a., 2019), we use average anchors to produce word translations to evaluate alignments. For the clustering based method, we use cluster-level average anchors of multi-sense words. Gold standard dictionaries are taken from the MUSE framework and contain 1,500 distinct source words.

### 5.2 Supervised Mapping

**Dictionary** The baseline supervised linear mapping is calculated based on a dictionary of 5,000 distinct source words downloaded from the MUSE library.
**Corpus for word occurrence embedding and anchor calculation** We compute the average of token embeddings on a fraction (around 500MB, or 80 million words) of English (/French) Wikipedia dumps as anchor vectors for the English (/French) words in dictionaries.

#### 5.2.1 Detailed Analysis about *bank*

To obtain an intuitive understanding of how multi-sense words behave in supervised mapping methods, we start our supervised mapping experiment focusing on a common English multi-sense word *bank*.
**2 dictionaries used for supervised linear mapping** To analyze the influence of incomplete translation pairs about *bank* in the dictionary, we generate two filtered dictionaries by removing translation pairs containing *bank* (form-based removal: *bank* $\Leftrightarrow$ *banques* and *bank* $\Leftrightarrow$ *banque*) and by removing translation pairs having the same lemma as *bank* (lemma-based removal: *bank* $\Leftrightarrow$ *banques*, *bank* $\Leftrightarrow$ *banque*, *banks* $\Leftrightarrow$ *banques*, *banking* $\Leftrightarrow$ *banques*, *banking* $\Leftrightarrow$ *banque*, and *banking* $\Leftrightarrow$ *bancaire*).

For token embeddings visualization, we compute token embeddings of the English word *bank* and of its French translations (i.e. "banque", "bord", "rive", and "berge", according to the Collins English-French Dictionary and WordReference.com) over around $500MB$ English and French corpora.

#### 5.2.2 Removal of English and (or) French Multi-sense Words

Based on the Wikipedia list of English homonyms, we generate two dictionaries by form-based removal and lemma-based removal. The original dictionary has $9496$ valid translation pairs, the form-based removal dictionary has $9161$ valid translation pairs and the lemma-based removal dictionary has $9076$.

For French, we generate four dictionaries by form-based removal and lemma-based removal based on two French polyseme lists. The form-based removal dictionaries have $9416$ and $9331$ valid translation pairs and the lemma-based removal dictionaries have $9370$ and $9226$ based on two lists respectively.

Furthermore, we also tried to remove both English and French Multi-sense Words by form-based removal and lemma-based removal.

### 5.3 Unsupervised Mapping

We calculate token embeddings for the 50,000 most frequent words in English and in the target language. For frequent words selection, we follow the word order in FastText pre-trained word vectors, which are sorted in descending order of frequency. The corpus used for anchor calculation and also the multi-sense word lists are the same as those used for supervised mapping.

To apply the spectral clustering algorithm to multi-sense word token embeddings, we calculate the frequency of token embeddings first. If it is less than 160, we keep the original average anchor embedding. If it is larger than 10,000, we randomly sample a subset of 10,000 token embeddings and then apply the clustering algorithm to it.

### 5.4 Set-up for Embedding Visualization

Embedding Projector[2] has been used for data visualization. We generate two 2-D graphs for each selected polysemy (or polysemies) by selecting PCA (Principal Component Analysis) for dimensionality reduction and Sphereize data *(The data is normalized by shifting each point by the [coordinates of the] centroid and making it unit [length])* for data normalization.

Note that PCA is approximate in the Embedding Projector, i.e., *for fast results, the data was sampled to 50,000 points and randomly projected down to 200 dimensions.* As token embeddings generated by ELMo have 1024 dimensions, the embeddings used for visualization were randomly projected down to 200 dimensions.

## 6 Results

| Alignment | 1st LSTM output layer | | | | | | 2nd LSTM output layer | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | nn | | | csls_knn_10 | | | nn | | | csls_knn_10 | | |
| | P@1 | P@5 | P@10 | P@1 | P@5 | P@10 | P@1 | P@5 | P@10 | P@1 | P@5 | P@10 |
| (a) Supervised Mapping | | | | | | | | | | | | |
| Baseline | **55.20** | 73.85 | 80.11 | 68.48 | 84.65 | 88.78 | **55.95** | **73.57** | 79.49 | **67.17** | 82.31 | 86.79 |
| Form-based removal (en) | 54.99 | 74.19 | 79.63 | **68.55** | **85.13** | 88.58 | 55.33 | 73.43 | 79.22 | 66.96 | **82.59** | 86.51 |
| Form-based removal (fr-1) | 54.85 | **74.26** | 79.77 | **68.55** | 84.86 | 88.92 | 55.88 | 73.50 | **79.63** | 66.90 | 82.11 | 86.79 |
| Form-based removal (fr-2) | 54.85 | 73.85 | **80.32** | 68.27 | 84.65 | 88.71 | 55.81 | **73.57** | **79.63** | 67.10 | 82.31 | 86.85 |
| Lemma-based removal (en) | 55.06 | 74.05 | 79.83 | 68.41 | 85.07 | 88.64 | 55.33 | 73.30 | 79.15 | 66.62 | 82.38 | 86.58 |
| Lemma-based removal (fr-1) | 54.92 | 74.19 | 79.83 | 68.07 | 84.79 | **89.13** | 55.82 | **73.57** | 79.56 | 66.83 | 82.17 | 86.79 |
| Lemma-based removal (fr-2) | 54.85 | 73.57 | 80.11 | 68.41 | 84.72 | 88.71 | 55.74 | **73.57** | **79.63** | 66.83 | 82.38 | **86.99** |
| (b) Unsupervised Mapping | | | | | | | | | | | | |
| Baseline | 42.81 | 62.70 | 67.72 | 48.11 | 69.99 | 74.54 | 35.58 | 49.90 | 56.64 | 42.60 | 62.42 | 68.62 |
| Anchors removal (en) | 52.44 | 67.38 | 72.06 | 57.88 | 73.43 | 77.22 | No convergence | | | | | |
| Anchors removal (fr-1) | 48.59 | 63.11 | 67.65 | 53.68 | 69.37 | 72.61 | **47.69** | **61.73** | **67.45** | **53.34** | **70.27** | **76.05** |
| Anchors removal (fr-2) | 45.97 | 60.16 | 64.30 | 50.52 | 65.33 | 69.81 | No convergence | | | | | |
| Anchors removal (en & fr-1) | No convergence | | | | | | 36.89 | 51.41 | 57.47 | 41.77 | 60.70 | 67.72 |
| Anchors removal (en & fr-2) | 51.96 | 68.44 | 73.12 | 58.17 | 75.53 | 79.53 | 33.43 | 45.83 | 50.59 | 39.83 | 53.55 | 59.27 |
| Anchors replacement (en) | **54.71** | **70.54** | **75.02** | **60.98** | **78.32** | **82.38** | No convergence | | | | | |

Table 3: Precision at $k = 1, 5, 10$ of bilingual lexicon induction from the aligned cross-lingual embeddings.

### 6.1 Visualization of the Token Embeddings of *bank*

Experiment results are shown in three figures presented below, in which dark blue points represent the English word *bank*, light blue points are token embeddings for the French word *banque*, and the French words *berge*, *bord*, *rive* are in green, red and pink colors respectively.

As shown in Figure 3, in the baseline aligned embedding space, the point cloud of *banque* is close to the middle part of the point cloud of *bank*. After removing the translation pairs containing words having the same form or lemma as *bank*, the point cloud of *banque* is moving to the top part of the *bank* point cloud, which is the cluster of the *financial institution* meaning of *bank*.

We take this as meaning that after removing incomplete supervision data (translation pairs in the dictionary) for multi-sense words, the alignment for multi-sense words is indirectly improved thanks to better supervision data for general embedding spaces mapping.

### 6.2 Lexicon Induction Task

In Table 3, we show the accuracy of the lexicon induction task based on different alignments.

[2]http://projector.tensorflow.org

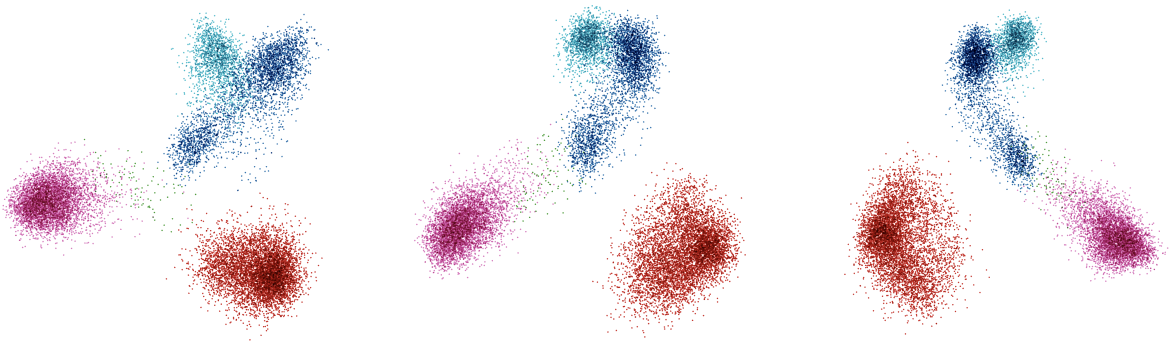

Figure 3: Aligned token embeddings for the English word *bank* (in dark blue) and French words *banque* (in light blue), *berge* (in green), *bord* (in red) and *rive* (in pink). Baseline alignment shown on the left, alignment after removing translation pairs having the same form as *bank* shown in the middle and alignment after removing translation pairs having the same lemma as *bank* shown on the right.

For supervised cross-lingual word embedding alignment, we found that removing translation pairs containing words having the same form or lemma as homonym words does not largely affect the lexicon induction task results (around $0.6\%$ difference in the precision at $k = 1$).

We observe below the difference between the baseline predictions and the form-based removal predictions (1st LSTM output layer, P@1) in two aspects:

- **Baseline prediction is correct while the form-based removal prediction is wrong.** In this case, we found some of the form-based removal predictions are indeed correct and that the gold standard is incomplete. For instance:
  1. Single-sense word: e.g., **highlight**, the predicted mapping of the form-based removal is *souligné*, but the gold standard is *souligne*
  2. Multi-sense word: e.g., **galaxy**, the predicted mapping of the form-based removal is French word *titan*, the gold standard is *galaxie, galaxy*, *galaxy* is a multi-sense word which has the meaning of a group of illustrious people; **commands**, the predicted mapping of the form-based removal is *instructions*, the gold standard is *commandements*, *commandes*. *instructions* is another meaning of English word *commands*.
- **Baseline prediction is wrong while the form-based removal prediction is right.** There are 11 words which are aligned correctly by the form-based removal, i.e, **flute**, gold standard is *flûte*, *flûtes*, the baseline method maps it to the French word *trompette*, which is another instrument trumpet; **madagascar**, the baseline prediction is *mozambique*, the name of one Africa country and also the Mozambique channel between Madagascar and the African mainland.

For unsupervised cross-lingual word embedding alignment (Table 3), we found that removing exact homonym-related anchor embeddings improves the P@top1 by 10 points and the P@top5 and P@top10 by 5 points (anchors removal(en)). Removing noisy information about multi-sense words is therefore very beneficial in this case. Replacing multi-sense word average anchor embeddings with cluster-level average anchors embeddings achieves the best result by using 1st LSTM output layer of ELMo.

## 7 Conclusion

In this paper, we explored the contextual word embeddings (token embeddings) of multi-sense words, argued that the current state-of-the-art method for cross-lingual token embedding learning cannot handle multi-sense words well and proposed our solutions by considering multi-sense word token embeddings as noise. Experiments showed that our methods can improve the token embeddings alignment for multi-sense words in a microscopic perspective without hurting the macroscopic performance on the bilingual lexicon induction task. As the research on cross-lingual token embedding learning is still in its early stage, we also discussed possible future work such as applying clustering algorithms on token embeddings to obtain sense-level multi-sense word representations.

Possible extensions would be to train a multi-sense word detector based on the number of clusters of token embeddings for each word and to create a new evaluation task for cross-lingual contextual word embeddings (token embeddings) with attention to multi-sense words.

## References


[Aldarmaki und Diab 2019] ALDARMAKI, Hanan ; DIAB, Mona: Context-Aware Cross-Lingual Mapping. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, Juni 2019, S. 3906–3911. – URL https://www.aclweb.org/anthology/N19-1391

[Conneau u. a. 2017] CONNEAU, Alexis ; LAMPLE, Guillaume ; RANZATO, Marc'Aurelio ; DENOYER, Ludovic ; JÉGOU, Hervé: Word translation without parallel data. In: *arXiv preprint arXiv:1710.04087* (2017)

[Devlin u. a. 2019] DEVLIN, Jacob ; CHANG, Ming-Wei ; LEE, Kenton ; TOUTANOVA, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, Juni 2019, S. 4171–4186. – URL https://www.aclweb.org/anthology/N19-1423

[Joulin u. a. 2017] JOULIN, Armand ; GRAVE, Edouard ; BOJANOWSKI, Piotr ; MIKOLOV, Tomas: Bag of Tricks for Efficient Text Classification. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain : Association for Computational Linguistics, April 2017, S. 427–431. – URL https://www.aclweb.org/anthology/E17-2068

[Lefever und Hoste 2009] LEFEVER, Els ; HOSTE, Veronique: Semeval-2010 task 3: Cross-lingual word sense disambiguation. In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions* Association for Computational Linguistics (Veranst.), 2009, S. 82–87

[Mikolov u. a. 2013a] MIKOLOV, Tomas ; CHEN, Kai ; CORRADO, Greg ; DEAN, Jeffrey: Efficient estimation of word representations in vector space. In: *arXiv preprint arXiv:1301.3781* (2013)

[Mikolov u. a. 2013b] MIKOLOV, Tomas ; LE, Quoc V. ; SUTSKEVER, Ilya: Exploiting similarities among languages for machine translation. In: *arXiv preprint arXiv:1309.4168* (2013)

[Peters u. a. 2018] PETERS, Matthew ; NEUMANN, Mark ; IYYER, Mohit ; GARDNER, Matt ; CLARK, Christopher ; LEE, Kenton ; ZETTLEMOYER, Luke: Deep Contextualized Word Representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana : Association for Computational Linguistics, Juni 2018, S. 2227–2237. – URL https://www.aclweb.org/anthology/N18-1202

[Ruder u. a. 2017] RUDER, Sebastian ; VULIĆ, Ivan ; SØGAARD, Anders: A survey of cross-lingual word embedding models. In: *arXiv preprint arXiv:1706.04902* (2017)

[Schuster u. a. 2019] SCHUSTER, Tal ; RAM, Ori ; BARZILAY, Regina ; GLOBERSON, Amir: Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, Juni 2019, S. 1599–1613. – URL https://www.aclweb.org/anthology/N19-1162

[Wang u. a. 2018] WANG, Quan ; DOWNEY, Carlton ; WAN, Li ; MANSFIELD, Philip A. ; MORENO, Ignacio L.: Speaker diarization with LSTM. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* IEEE (Veranst.), 2018, S. 5239–5243

[Xing u. a. 2015] XING, Chao ; WANG, Dong ; LIU, Chao ; LIN, Yiye: 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*, 2015, S. 1006–1011

[Yang u. a. 2019] YANG, Zhilin ; DAI, Zihang ; YANG, Yiming ; CARBONELL, Jaime ; SALAKHUTDINOV, Ruslan ; LE, Quoc V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: *arXiv preprint arXiv:1906.08237* (2019)

# Appendix

| Word | Cluster Positions | Sentences @ 1st layer | Sentences @ 2nd layer |
|---|---|---|---|
| bank | left | Small Craft Company USMC assisted in locating the bodies of the slain snipers and were engaged in a large fire fight on the east **bank** of the Euphrates River in the city of Haditha. | At the northern **bank** of the Svir River () the Finnish army had prepared a defence in depth area which was fortified with strong-points with concrete pillboxes, barbed wire, obstacles and trenches. |
| | | The population on the east **bank** of the Weser had not prepared adequate defenses, so the crusading army attacked there first, massacring most of the population; the few survivors were burnt at the stake. | These specimens were collected at the Karagachka locality (locality 34 or PIN 2973), to the opposite **bank** of the Karagatschka River from Karagachka village located in a drainage basin of left bank of the Ural River, Solâ Iletsk district of Orenburg Region, southern European Russia. |
| | right | If government bonds that have come due are held by the central **bank**, the central bank will return any funds paid to it back to the treasury. | Issue **bank** notes; |
| | | Liz is astonished when the police suddenly arrive at the pub to tell her that Jim has been caught robbing a **bank** and now has a number of hostages. | Although such measures were not effected, the new administration was successful in tackling other issues: both deficit and the cost of living dropped while the **bank** reserves trebled, and some palliatives were introduced in lieu of a land reform (the promised tax cuts, plus the freeing of "mainmorte" property). |
| spring | top left | However, after reaching Ulster the horse stops and urinates, and a **spring** rises from the spot. | The **spring** had been shut off by a rock 74 meters long and 30 meters wide, which obstructed the construction of a running water system. |
| | | Over running water â Literally "living", that is, **spring** water. | The holy **spring** is known to change its colour with various hues of red, pink, orange, green, blue, white, etc. |
| | bottom left | A 5'10", 170-pound infielder, Werber was at **spring** training and toured for several weeks in July with the Yankees in 1927. | Joss attended **spring** training with Cleveland before the start of the 1911 season. |
| | | He was invited to **spring** training and sent to minor league camp on March 14. | He pitched in the California Angels minor league system in the early 1990s and participated in "Replacement player" **spring** training games in 1995 for the Toronto Blue Jays. |
| | bottom middle | In **spring** 912, the Jin attack against Yan got underway, with Zhou commanding the Jin army in a joint operation with the Zhao general Wang Deming (Wang Rong's adoptive son) and the Yiwu Circuit (headquartered in modern Baoding, Hebei) army commanded by Cheng Yan (whose military governor, Wang Chuzhi, was also a Jin ally). | In **spring** 2017, Ponders hit the road supporting Pouya and Fat Nick, opening to sellout crowds across Ontario and Quebec. |
| | | In **spring** 2010 CSX railroad removed the diamonds connecting the southern portion of the Belt Railroad, thus isolating the line from the U.S. rail system. | In **spring** 1944, the Rabstejn sub-camp of Flossenburg was created here, with a capacity of 600 prisoners. |
| | right | In the **spring** of 1935, the All-Union Organization of Cultural Relations with Foreign Countries agreed to send a delegation to the upcoming First International Festival of the Folk Dance in London. | Hirsig's role as Crowley's initiatrix reached a pinnacle in the **spring** of 1921 when she presided over his attainment of the grade of Ipsissimus, the only witness to the event. |
| | | In the **spring** of 2012 in Pakistan was established Pakistani mission. | Brown wrote, "In the **spring** of 1819 a nightingale had built her nest near my house. |
| check | left | Because the defined cases are exhaustive, the compiler can **check** that all cases are handled in a pattern match: | It is standardized for use by mud engineers to **check** the quality of drilling mud. |
| | | Most spotters maintained books of different aircraft fleets and would underline or **check** each aircraft seen. | The lowest level, where the sounds are the most fundamental, a machine would **check** for simple and more probabilistic rules of what sound should represent. |
| | right | Usually, the trial **check** will quickly reject the trial match. | It is important to realize that glucose-glutamic acid is not intended to be an accuracy **check** in the test. |
| | | The donor's hematocrit or hemoglobin level is tested to make sure that the loss of blood will not make them anemic, and this **check** is the most common reason that a donor is ineligible. | U.S. Attorney General John Mitchell, citing an extensive background **check** by the Justice Department, was willing to forgive, stating that it was unfair to criticize Carswell for "political remarks made 22 years ago." |
| clear | top | From here, she had to fight an uphill battle to **clear** her name and proved her right by finding the authentic painting, while she was also struggling with financial hardship and interference from Min Jung-hak. | On 1 November, Ouagadougou Mayor Simon CompaorÃ© led volunteers on "Operation Mana Mana" (Operation Clean-Clean in Dyula) to **clear** the streets, which earned him praise on social media. |
| | | Jones' shoulder injury came after Botha attempted to **clear** him from a ruck and the Bulls star was subsequently cited and banned for two weeks for the challenge. | Again a gold medal favourite in the 110 metre hurdles at the London Olympics he pulled his Achilles tendon attempting to **clear** the first hurdle in the heats. |
| | Bottom left | She made it **clear** that she did not intend for Nassar to ever be free again. | Hugenberg for his part regarded "Katastrophenpolitik" as a good idea that was unfortunately abandoned, and made it **clear** that he wanted a return to "Katastrophenpolitik". |
| | | Many Southerners felt that the Compromise of 1850 had been shaped more towards Northern interests; the Georgia Platform made it **clear** that the future of the nation depended on the North strictly adhering to the Compromise. | The political heat was turned up on the issue since Bush mentioned changing Social Security during the 2004 elections, and since he made it **clear** in his nationally televised January 2005 speech that he intended to work to partially privatize the system during his second term. |
| | Bottom right | However, in "Reference re Secession of Quebec", the Supreme Court of Canada has essentially said that a democratic vote in itself would have no legal effect, since the secession of a province in Canada would only be constitutionally valid after a negotiation between the federal government and the provincial government; whose people would have clearly expressed, by a **clear** majority, that it no longer wished to be part of Canada. | He was the **clear** winner with ten seconds over the runner-up, fellow Kenyan Albert Kiptoo Yator. |
| | | The game sees Kasparov rejecting **clear** drawing opportunities and eventually losing. | He wrote to Irene Tasker in South Africa, in a **clear** hand, telling her how much better he was. |

Table 4: Corresponding sentences selected from the token embedding clusters of the English words *bank*, *spring*, *check* and *clear*.