Unsupervised Word Translation for Fine-Grained Professional Terms

Chen Almagor
School of Computer Science and Engineering
The Hebrew University of Jerusalem
chen.almagor@mail.huji.ac.il

Abstract

Unsupervised word translation works reasonably well at the word level, but it not necessarily accurate at the fine-grained professional terms level.
In this work, we aimed to deal with this challenge by examining two approaches: (1) Projecting professional terms to the global language space, and utilizing the received embeddings for matching or mapping between the terms. (2) Using a known global translation between languages, to learn a specialized translation function for particular domain.
Success in the area of professional terms translation can leverage machine translations accuracy, as well as contributing the challenge of rare words translation.
None of the explored models in this work have shown a signal for improving the global transformation performance. Though, this work contributes for the understating of future directions for this task.

1 Introduction

Inferring word translations between languages is a long-standing research task. Following the success of monolingual word embeddings, a number of studies have explored cross-lingual word embeddings (Ruder et al. (2019)). The goal is to learn word vectors such that similar words have similar vector representations regardless of their language.
Commonly, cross-lingual word embedding methods are based on a mapping approach, which have shown to be a very effective way (Mikolov et al. (2013)). The underlying idea is to independently train the embeddings in different languages using monolingual corpora, and then map them to a shared space through a linear transformation. The map can then be used to translate words between the language pair.
There is line of works that managed to implement this mapping approach method in an unsupervised manner, based on distributional information alone (Conneau et al. (2017); Artetxe et al. (2018); Hoshen and Wolf (2018)). They learn an initial transformation that bootstrapping the process without any parallel words groundtruth, and refining the resulting transformation with unsupervised criterion. The initialization is a core component of any mapping-based approach, and in unsupervised approaches in particular. While Conneau et al. (2017) learn this initial mapping in an adversarial way (Goodfellow et al. (2014)), Hoshen and Wolf (2018) introduce a simple and non-adversarial method, and use an Iterative Closest Point (ICP) approach initialized with a randomized PCA.

Those methods achieve a high accuracy in the global words translation, but they are less accurate at the fine-grained professional terms level. For example, Hoshen and Wolf (2018) method achieves 81.1% accuracy on global English to Spanish word translation, while on professional medical terms it achieves 54.41%.

In this work, we aimed to deal with this challenge by examining two approaches:

Preprint. Under review.

1. Projecting professional terms to the global language space, and utilizing the received embedding for matching or mapping between the terms.

2. Using a known global translation between languages, to learn a specialized translation function for particular domain.

For the first approach, we used two independent dictionaries of a professional terms and their definitions, and utilized a multilingual shared encoder to embed the definitions. The received embedding were matched by a reciprocal nearest neighbors, resulting an unsupervised inferred lexicon. Although it has lower accuracy compared to the global translation, this lexicon is proved as a stable initialization for the unsupervised mapping, meaning the model convergences in every time it was initialized with it.
Another variation for this approach was to align before retrieving the reciprocal nearest neighbor, but it observed as less effective.

For the second approach, we chose the non-adversarial unsupervised approach, and used Hoshen and Wolf (2018) as the base for our methods. We assumed that the professional in-domain terms are a subspace of the global space, and explore the influence of initializing and training by in-domain data. During the research we observed that the linear solution seems to reach its limit, and we start to explore if there exists a non-linear variation which can be based on this regression approach.
Nakashole and Flauger (2018) claims that the underlying map is expected to be non-linear but in small enough neighborhoods can be approximated by linear maps. Following this approach, we proposed a clustering variation for the base method. We first approximate the mapping on the whole space, then cluster the source and the target spaces, and learn a mapping between the clusters, utilizing the learned mapping for matching the clusters and for initializing of the clusters specific mappings.

The suggested methods are unsupervised, but in order to isolate the unsupervision effect on the results, we implemented a supervised variations, which helps us to conclude and analyze the results better.

None of the examined models have shown a signal for improving the global transformation results. Though, we can conclude regarding the future directions for handling this challenge.

During our research, we focused on a specialized domain: medicine, and on English and Spanish languages.

2 Previous Works

2.1 Unsupervised word translation

Cross-lingual embedding mappings have shown to be an effective way to learn bilingual word embeddings (Mikolov et al. (2013)). These methods work as follows: for a given pair of languages, first, monolingual word vectors are learned independently for each language, and second, under the assumption that word vector spaces exhibit comparable structure across languages, a linear mapping function is learned to connect the two monolingual vector spaces. The map can then be used to translate words between the language pair.
These mapping methods can be approached in a supervised manner, by using a bilingual dictionary of a few thousand entries to learn the mapping (Mikolov et al. (2013)), in semi-supervised manner, where the training dictionary is much smaller and used as part of a bootstrapping process (Artetxe et al. (2017)), and in a fully unsupervised manner, which attempted to learn cross-lingual embedding mappings based on distributional information alone (Conneau et al. (2017); Artetxe et al. (2018); Hoshen and Wolf (2018)).
The initialization is a core component of any mapping-based approach, and in unsupervised approaches in particular, which requires to learn an initial transformation that bootstrapping the process without any parallel inter-lingual corpora. Conneau et al. (2017) learn an initial mapping leveraging adversarial training (Goodfellow et al. (2014) by additionally training a discriminator to differentiate between projected and actual target language embeddings. Artetxe et al. (2018) propose to use an initialization method based on the heuristic that translations have similar similarity distributions across languages. Hoshen and Wolf (2018) first project vectors of the N most frequent words to a lower dimensional space with PCA, and use an ICP approach to find a transformation that minimize the sum of Euclidean distances and enforce cyclical consistency constraints.

Note that unsupervised methods greatly benefit from a refinement procedure, often based on some variants of Iterative Closest Points (ICP, Besl and McKay (1992)). It explains the reason that Hoshen and Wolf (2018) approach, that relying on ICP, was able to achieve great performance, with a simple and non-adverserial method.

Although those unsupervised word translation methods gain high accuracy for the global language translation, a lower accuracy was observed on a professional medical terms. In this work, we attempted to find an adjustment of Hoshen and Wolf (2018) method to the professional terms distribution. In addition, we infer a professional terms lexicon in an unsupervised way and utilize it as an initialization for this method.

Hoshen and Wolf (2018) method is described in detail in section 2.1.1

2.1.1 Mini-Batch Cycle ICP

Our work using Mini-Batch Cycle ICP (MBC-ICP), introduced by Hoshen and Wolf (2018), as the base global translation method. Therefore, following is a dedicated review of it.

Mini-Batch Cycle ICP (MBC-ICP) is a non-adversarial unsupervised word translation method, which is a modified version Iterative Closest Point (ICP). ICP is a popular optimization method for minimizing Equation 1.

MBC-ICP learns transformations T_{xy} for $X \rightarrow Y$ and T_{yx} for $Y \rightarrow X$. Each iteration of this method computes the nearest neighbor matches from the transformed word, and then optimize transformations T_{yx} and T_{xy} by minimizing the distances of each mapped words and enforcing a cyclical consistency constraints, which force vectors round-projected to the other language space and back to remain unchanged.

This method composed of three steps:

1. *Initialization* - PCA-MBC-ICP - For each language $[X, Y]$, select the N most frequent word vectors. Project the word embedding after centering, to the top p principle components, and find the best initialization permutation by running iterations of MBC-ICP.
2. *Training* - RAW-MBC-ICP - Run iterations of MBC-ICP on the original word vectors (no PCA).
3. *Training on Reciprocal Pairs* - Run RAW-MBC-ICP only on pairs that are likely to be correct matches. Those pairs can be identified using the transformations from the previous step, and by applying the reciprocity heuristic: if a pair of words is matched in both $X \rightarrow Y$ and $Y \rightarrow X$ directions, the pair is denoted reciprocal.

The output of the method are transformation matrices T_{xy} and T_{yx}.

2.2 Non-linear mapping

Mapping-based approaches assume that a linear transformation can project the embedding space of one language into the space of a target language. In practice, Mikolov et al. (2013) obtained better results on the word translation task using a simple linear mapping, and did not observe any improvement when using more advanced strategies like multilayer neural networks.

However, Nakashole and Flauger (2018) claims that the underlying maps are non-linear, and demonstrate it by locally approximate these maps using linear maps, and show that they vary across the word embedding space.

According to this finding, we suggest a clustering variation for Hoshen and Wolf (2018).

2.3 Clinical Professional-Consumer Languages Translation

In the area of medical terms, there is a research effort to translate between professional and consumers languages, meaning to simplify the clinical language in order to improve patient-clinician communication. As part of those works, Weng and Szolovits (2018) and Weng et al. (2019) utilized the embeddings alignment methods for mapping between unparalleled clinical professional and consumer language embeddings. They first independently trained word embeddings on both the

corpus with abundant clinical professional terms and the other with mainly health care consumer terms. Then, they aligned the embeddings, investigating both Conneau et al. (2017) and Artetxe et al. (2018) implementations. The identical strings served as anchors for initialization, and adversarial learning were adopted if identical strings were not used.
Another work that deals with the relation between the professional language and the global language is Limsopatham and Collier (2015). They propose to adapt an existing phrase-based machine translation technique and a vector representation of words to map between a social media phrase and a medical concept.
Those works were the inspiration for the inferred unsupervised lexicon method, in the idea of transferring the professional term to a general simple language, which has a translation solutions.

2.4 Cross-Lingual Sentence Encoding

There has been a growing interest in cross-lingual language understanding and transferring, and an effort on developing multilingual sentence embeddings (Schwenk and Douze (2017); Hermann and Blunsom (2014); España-Bonet et al. (2017)).
Yang et al. (2019) introduced a multilingual models for embedding sentence-length text into a shared semantic embedding space. We employed this encoder to embed the professional terms definitions. Important to note that they augment unsupervised learning with training on supervised data, but there exists a fully unsupervised cross-lingual language models that can be adopted (Lample and Conneau (2019), for example). Those cross-lingual sentences representations are utilized for semantic retrieval applications, in particular for machine translation.

3 Methods

In the following section we describe our unsupervised methods, which aimed to improve the professional terms translation.

Let us define two languages X and Y, each containing a set of N_X and N_Y professional terms, $x_1, ...x_{N_X}$ and $y_1, ...y_{N_Y}$ respectively. We denote by $f(n)$ the correspondence function such that for every x_n, $f(n)$ yields the index of the Y term that corresponds to the term x_n.

When we refer to an alignment transformation, the goal is to find a transformation, T, that will align every x_i from language X to its corresponded $y_{f(i)}$ in language Y. The objective is therefore to minimize:

$$\underset{T}{\operatorname{argmin}} \sum_i \min_{f(i)} |y_i - T(x_{f(i)})| \tag{1}$$

3.1 Unsupervised Inferred Lexicon

Inspired by the translation between clinical professional terms to simple language terms (Weng et al. (2019), we wanted to build a mechanism that project a professional term into the global language space, and retrieve the translation from it. The idea is to bypass the unknown distribution and structure of the professional terms domain, transfer it to a known distribution space, and utilize known methods for translating sentences to tackle our unsolved task.

To implement this approach, we chose to use an in-domain dictionaries, that composed of professional terms and their definitions. The definitions are sentences in a global language, while the terms are professional. Note that those dictionaries are independent, meaning that the source and the target dictionaries may includes different terms.
We utilize a pre-trained retrieval focused multilingual sentence encoder (Yang et al. (2019), which embed text from multiple languages into a single semantic space, and encode a fixed-size vector representation for each definition.

We denote the encoder transformation by E, and the corresponding definitions by $x_1^{def}, ...x_{N_X}^{def}$ and $y_1^{def}, ...y_{N_Y}^{def}$. The encoder embeddings are therefore defined by $E(x_1^{def}), ...E(x_{N_X}^{def})$ and $E(y_1^{def}), ...E(y_{N_Y}^{def})$.

3.1.1 Inferring according to the definitions similarity

For each definition, we retrieve the most semantic similar definition by choosing its nearest neighbor from the other language, and infer the lexicon by the reciprocal definitions pairs.
Formally - for every definition $y^{def} \in Y^{def}$, we find the nearest definition to $E(y^{def})$ from the set $\{E(x^{def}) \mid x^{def} \in X^{def}\}$. The same is done for the other direction. If a pair of definition is matched in both $X^{def} \rightarrow Y^{def}$ and $Y^{def} \rightarrow X^{def}$ directions, the pair is denoted reciprocal, and the corresponded words of those definitions are matched.

3.1.2 Inferring according to alignment of the definitions

In this case, we attempt to find a transformation, T, that will align every $E(x_i^{def})$ from language X to a its corresponded $E(y_{f(i)}^{def})$ in language Y. The objective in this case is to minimize:

$$\underset{T}{\operatorname{argmin}} \sum_i \min_{f(i)} |E(y_i^{def}) - T(E(x_{f(i)}^{def})|$$

We learn the alignment by Mini-Batch Cycle ICP (MBC-ICP) method, but initializing with the inferred lexicon that was described previously (instead of the original PCA-MBC-ICP initialization). We infer the lexicon by the reciprocal pairs in the end of the training.
Elaboration on the alignment method can be found in section 3.2.

3.2 Variations of Mini-Batch Cycle ICP

We aimed to utilize a known global translation method between languages, to learn a specialized translation function for particular domain. We chose the non-adversarial unsupervised approach, used Mini-Batch Cycle ICP as the base method, and tried to adjust it for the professional terms translation task.

3.2.1 Using In-Domain Data

The original MBC-ICP method uses the most frequent words in each language for the initialization and training.
In order to explore the professional terms distribution, we used professional terms data (medical terms, in our case).
Though, no improvement was observed by only using an in-domain data.

An important note is that when applying the original method on professional terms instead of the most frequent words of each language, it not necessarily convergence with the original setup. A larger amount of terms was required for the initialization, as well as search for more stochastic solutions at the random restarts.

3.2.2 Utilizing the Inferred Lexicon for Initialization

Naturally, iterative approach relies on a good initialization. Therefore, we attempted to improve this step.
We modified the initialization of the original MBC-ICP by using the inferred lexicon from section 3.1.1. Note that we did not observe any improvement by applying the original PCA-MBC-ICP procedure on the inferred lexicon words (instead of the identity function).

This method is proved as more stable in terms of the convergence of the model, but is not contributed for improving the performance of professional terms translation.

3.2.3 Clustering and Aligning Per Cluster

Nakashole and Flauger (2018) investigated the behavior of mappings between word embedding spaces of different languages. They claim that the underlying map is expected to be non-linear, but in small enough neighborhoods can be approximated by linear maps.

Attempting to improve the global transformation, and in order to investigate if the distribution in the professional terms domain behaves as described above, we examine another modification for the original mapping method.

This method works as follow:

1. Run MBC-IPC linear mapping method on professional terms, which output two transformations, T_{xy} and T_{yx}, as well as the reciprocal pairs in the end of the training, denoted by RP.
2. Divide the professional terms $x \in X$ into k clusters $\{C_1^X, ...C_k^X\}$, as well as $y \in Y$ terms: $\{C_1^Y, ...C_k^Y\}$.
3. Match the clusters, utilizing the reciprocal pairs from step 1 (see details below). We denote by $g(j)$ the index of the matched cluster of C_j^X.
4. Initialize each $T_{C_j^X C_{g(j)}^Y}$ and $T_{C_{g(j)}^Y C_j^X}$ with T_{xy} and T_{yx}, respectively.
5. Run MBC-IPC on each clusters pair separately, resulting $2k$ optional transformations $\{(T_{C_j^X C_{g(j)}^Y}, T_{C_{g(j)}^Y C_j^X}) \mid j = 1, ..k\}$

In inference time, search for which of the clusters the source word belongs, and apply the relevant transformation on it.

Unsupervised Clusters Matching: Due to the unsupervised setup, matching the clusters is required. We use the reciprocal pairs, from the end of the training of step 1 to determine which clusters are related.

C_j^X is matched to $C_{g_x(j)}^Y$ if it maximize:

$$g_x(j) = \max_{t=1,..k} \sum_i \begin{cases} 1 & x_i \in C_j^X \wedge y_{f(i)} \in C_t^Y \wedge (x_i, y_{f(i)}) \in RP \\ 0 & otherwise \end{cases}$$

The same process is formed also in the other direction, to match C_j^Y and $C_{g_y(j)}^X$, for every j.
The final match is the reciprocal clusters pairs (ideally).

The clusters matching is not necessarily one-to-one and onto in each direction as well as not necessarily fully reciprocal.
First, this case can imply that the amount of the chosen clusters did not represent the departures correctly. To examine if this indeed the issue, we experimented a range of clusters amount.
Second, it can be caused if the reciprocal pairs are not divided across all the clusters. In this case we can use the transformations from step 1, transform the source to the target, and use the same matching method that described above (meaning, we remove the reciprocal constrain on the original mapped words).

4 Experiments

In the following section, we describe the our research experiments.

We performed our experiments on the medical domain, and on English and Spanish languages.

4.1 Data

We employed `MedicineNet MedTerms Medical Dictionary` [1] as the English medical dictionary, which contains 16,522 terms. For the Spanish dictionary, we used `Clinica Universidad de Navarra`[2] medical dictionary, which contains 17,793 terms.
We utilized those dictionaries both for the inferred lexicon methods, that requires a professional terms dictionaries (of terms and their definition), and as an unsupervised medical vocabularies.
Note that the definitions were not parsed or manipulated, in order to contain as much information as possible.

[1] https://www.medicinenet.com/medterms-medical-dictionary/article.htm

[2] https://www.cun.es/diccionario-medico

For evaluation, we used `123 Teach Me, Spanish for Medical Professionals`[3] glossary. This glossary contains both the professional and the simple translation for every term. We extracted the professional translations, and got $2,067$ professional words lexicon (without phrases). Another parallel glossary that we employed is `MedSpEN`[4] medical English-Spanish glossary (Villegas et al. (2018). `MedSpEN` contains $34,033$ words, and $110,678$ phrases. This glossary is utilized both for evaluation and for the supervised methods implementation.

4.2 Supervised Setups for Investigation

In order to isolate the effect of the unsupervised setup, and to understand the behaviour of the professional medical terms domain, we implemented a supervised versions of the proposed methods.

The supervised models are mentioned along the experiments section, as a base-line for comparison and conclusions about the suggested unsupervised models.

`MedSpEN` glossary, which contains parallel pairs of professional medical terms and their translations, was utilized both for evaluation and for the supervised methods implementation. While using it for the training - it was divided to training and test sets (unless it was an overfitting experiment).

MBC-ICP supervised setup We used the supervision for initializing with the correct matches, and training corpora that contain the same words. Note that the optimization is still done on the original unsupervised loss. The supervision is reflected by the initialization (perfect anchors), and by the fact that for each word there exists a match in the other space.

Clustering supervised setup The clustering variation composed of two components that can be implemented in a supervised manner: (1) *Running the MBC-ICP* - the supervision was applied in the same way that described above, both when running on the whole space, and when running in the clusters level. (2) *Matching the Clusters* - in a supervised setup, the clusters of the target language was set according to the clusters of the source language, meaning - no match was needed. In unsupervised setup, the spaces are clustered independently, resulting in a varied sizes and not perfectly populated clusters. Therefore, and a matching between the clusters is required.

Neural Network supervised setup This regression task was converted to a simple neural network, with fully supervision setup. In these experiments the network was trained on parallel words embedding, meaning the loss computed with the supervised translation of each word. The idea was to compare between linear and non-linear functions, without any unsupervised concern, in order to have a conclusion regarding the relevant model for this task.

4.3 Implementation Details

Except for the inferred lexicon method, we applied FastText (Bojanowski et al. (2016)) for the word embeddings, which uses the internal word co-occurrence statistics for each language. We used a model that was trained on Wikipedia corpora, which embed the words to 300 dimensional vectors.

4.3.1 Inferred Lexicon

We used a multilingual universal sentence encoder for semantic retrieval (Yang et al. (2019)). This model is based on convolutional neural network (Kim (2014), and embed text from 16 languages into a single semantic space, showing strong performance on cross-lingual retrieval. The input is variable length text and the output is a 512 dimensional vector.
The semantic search engine was implemented by `SimpleNeighbors`[5] library, which is a wrapper for `Annoy`[6] library, to efficiently look up results from the corpus. The index was built on top of dot (inner) product metric.

[3] https://www.123teachme.com/medical_dictionary

[4] https://github.com/PlanTL-SANIDAD/MeSpEn_Glossaries

[5] https://github.com/aparrish/simpleneighbors

[6] https://github.com/spotify/annoy

For learning transformation on top of the encoder embedding, we initialize with the inferred lexicon of the reciprocal definitions pairs, and applied MBC-ICP with the same setup that is described in section 4.3.2. The lexicon is inferred by the reciprocal words in the end of the training.

4.3.2 MBC-ICP Variations

For the MBC-ICP method we used the same parameters as the global method implementation[7]. Note that we did not observe an improvement by trying to increase the epochs or change the hyper-parameters.
We optimized the euclidean distance, and used ADAM optimizer, with learning rate of e^{-2}.
The initialization were varied across the experiments, see 4.4 section more details.

The clustering was implemented by `FAISS` [8] library. The `scikit-learn (sklearn)` toolkit was tried as well, but the results were similar.

4.3.3 Neural Network

To implement the neural network model we used `Keras`[9].

The linear network was implemented by one fully connected layer. The non-linear network was implemented by one fully connected layer with varied activation (we examined ReLU, Leaky ReLU, Sigmoid and Tanh).

Both of the models were optimized by ADAM optimizer with learning rate of e^{-2}, and mean squared error (MSE) loss function. They run for 80 epochs.

4.4 Evaluation and Results

The evaluation concentrated on the word translation accuracy, measured by the fraction of words translated to a correct meaning in the target language (precision@k).
The similarity between the words calculated by Cross-domain Similarity Local Scaling (CSLS, Conneau et al. (2017)).
The evaluation was done using FastText embeddings, trained on Wikipedia of 300 dimensions.

4.4.1 Inferred Lexicons Results

By inferring the lexicon from the reciprocal definitions similarity, we received $2,689$ matches, 854 of them are words (not phrases).
We compared the accuracy of the words lexicon with the global method translation (on the same words). The accuracy comparison can be seen in Table 1.

Table 1: Accuracy (%) comparison between the global MBC-ICP and the inferred lexicon by medical definitions similarity

	P@1		P@5		P@10	
	en-es	es-en	en-es	es-en	en-es	es-en
Inferred Lexicon	44.94	35.15	56.27	45.05	62.36	50.68
Global MBC-ICP	87.10	86.86	93.55	93.51	95.29	96.24

We analyzed the results and attempted to characterize if there are kind of words in which the inferred lexicon success better.
We calculated the cosine similarity of each proposed translation from the most similar translation option, and sorted the results according to the Levinstein distance between them. We noticed that there is a drop in the average cosine similarity of the global model translations with low similarity

[7] `https://github.com/facebookresearch/Non-adversarialTranslation`

[8] `https://github.com/facebookresearch/faiss`

[9] `https://github.com/keras-team/keras`

by Levenshtein distance (from 0.67 to 0.5, for $similarity < 0.25$). The inferred lexicon had 0.6 average cosine similarity in this case.
Therefore, we extracted those words and check the accuracy again.
The global translation model still got a better results. Those results can be seen in Table 2.

Table 2: Accuracy (%) comparison on words with low similarity by Levenshtein distance (English-Spanish)

	Total	Levenshtein similarity < 0.25			
	Consine Similarity*	Consine Similarity*	P@1	P@5	p@10
Inferred Lexicon	0.67	0.60	49.25	61.19	68.65
Global MBC-ICP	0.68	0.50	86.56	94.02	95.52

*Average of the cosine of each proposed translation from the most similar translation option.

The inferred lexicon from the learned alignment of the definitions contains $3,202$ matches (with phrases) and $1,036$ words. The accuracy results can be seen in Table 3.

Table 3: Accuracy (%) comparison between the global MBC-ICP and the inferred lexicon by medical definitions alignment (English-Spanish)

	P@1	P@5	P@10
Inferred Lexicon	29.89	40.36	45.52
Global MBC-ICP	62.21	75.11	80.72

Combining the results above, we can conclude that the encoded definitions embeddings are not an accurate representation for the professional terms. Though, those lexicons contains matches that are not perfectly accurate, but they are close to the required translation. For example, ‘myalgia‘ (muscle pain), was matched to ‘algospasmo‘(cramp or painful muscle spasm), while the required translation is ‘mialgia‘. Therefore, this lexicon can be utilize as a stable initialization for other methods.

4.4.2 MBC-ICP Linear Variations Results

Those experiments were evaluated on English to Spanish model.

During the research, we performed large amount of experiments, both on the unsupervised and the supervised setups, in order to find the signal for improvement. Those experiments examined the data level, the initialization procedure, and the hyper-parameters.
In the data level, we sampled from the training data and attempted to find a subset that improving the final accuracy; We sorted each corpus by frequency and trained on the top 10K frequent words; We trained on a upper case terms embeddings, and on lower case terms embeddings; We trained at words level, and at pharse level.
In the initialization level, we applied the original PCA-MBC-ICP procedure on top of the initialization lexicon (both for the supervised and the unsupervised), as well as initialize with those lexicons at identity function.
In the hyper-parameters levels, we increased the epochs amount, and changed the optimizer learning rate to e^{-4}.

We also tested the results on different test sets, as well as on different portions of MedSpEN glossary.

Though, no dramatic improvement was observed between the global word translation model and the proposed models.

We chose a representative results from the unsupervised setup, as well as interesting results from the supervised setups (results that leaded us to conclusions), and presented them in Table 4.

Following is a description of the setups that we present in Table 4:

- **Unsupervised** - *Initializing* with the $2,690$ terms of the inferred lexicon by definitions similarity (identity), *training* on those terms, and *training on reciprocal pairs* of the independent dictionaries terms ($16,522$ terms, sampled from the Spanish corpus).
  Note that reciprocal pairs training on $7,500$ words got 51.75% on P@1.
- **Unsupervised with semi-supervised reciprocal pairs training** - *Initializing* and *training* as in the fully unsupervised setup, *training on reciprocal pairs* of the concatenation of the the unsupervised and the supervised corpora.
  With this setup we wanted to examine if the increase in the accuracy with the supervised setup is a result of a better initialization. This model achieved similar results to the supervised model, meaning that the improvement is because of better reciprocal options, and not because of a lower initialization quality.
- **Supervised** - *Initialization*, *training* and *reciprocal pairs training* on `MedSpEN` corpora ($20,812$ words, duplicated words were removed).
  This setup was also evaluated on the same data, to examine if it can be overfitted. The results are close to the global model performance, which provided the understanding that the linear model reach its limit, and a non-linear approaches should be taken in this case.
- **Supervised with less frequent terms** - *Initialization*, *training* and *training on reciprocal pairs* on `MedSpEN` corpora, but starting from the 5000 most frequent word.
  We noticed that the most frequent words are related to the global languages (for example - `‘health’, ‘body’, ‘support’`). In order to examine their influence, we removed them and started from the $5,000$ most frequent word (meaning that the model was trained mostly on professional terms).
  The training with the frequent words gained higher accuracy. We can conclude that those words are probably helps the model to generalize and infer the language model better.

Table 4: Accuracy (%) comparison between the global MBC-ICP and the examined variations setups

	Spanish for Medical			MedSpEN		
	P@1	P@5	P@10	P@1	P@5	P@10
Global MBC-ICP	61.51	77.25	81.39	54.41	68.91	72.29
Unsupervised	60.88	75.70	79.26	52.72	62.36	50.68
Unsupervised with semi-supervised reciprocal pairs training	65.02	78.86	81.96	-	-	-
Supervised	65.07	79.14	81.96	*59.29	*70.22	*73.18
Supervised with less frequent terms	61.63	76.39	80.41	-	-	-

* Overfitting attempt results

4.4.3 MBC-ICP Clustering Variation Results

Using the unsupervised configuration got poor accuracy (both for the fully reciprocal clustering match and for the match by the whole space alignment).
We suspected that the cause is the lack of samples in the unsupervised corpus, as the whole space unsupervised transformation on $7,500$ professional terms got lower accuracy. Therefore, we used the semi-supervised corpus, a concatenation of the unsupervised and the supervised professional terms, for the reciprocal pairs training of the whole space step, and for the clustering step. The results were better, but did not improve the whole space linear transformation.
Examining the supervised setup results, we can conclude that each cluster should have a sufficient amount of samples in order to convergence and to have the ability to generalize. It may explain the drop in the results from 7 clusters to 15 and 43 clusters. On the other hand - using small amount of clusters resulting similar accuracy as the non-clustered training.

The results can be seen on Table 5 and Table 6.

Table 5: Accuracy (%) comparison between clustering method setups (evaluated on Spanish for Medical glossary)

Clusters	P@1	P@5	P@10
Fully unsupervised setups			
No clustering	60.88	75.70	79.26
3*	44.97	62.43	67.14
3	45.77	63.06	67.94
5*	24.23	39.40	44.97
5	20.21	35.55	40.78
7	5.57	11.60	16.08
Unsupervised matching with semi-supervised data			
No clustering	65.02	78.86	81.96
3	43.02	61.17	67.03
5	51.12	69.04	73.63
7	64.84	78.86	82.36
Fully supervised setups			
No clustering	65.07	79.14	81.96
3	65.82	79.26	82.53
7	67.71	79.72	82.53
15	60.19	75.47	78.4
43	15.56	23.31	26.19

* With reciprocal clusters constraint

Table 6: Overfitting attempts - Accuracy (%) comparison between supervised clustering method to the supervised non-clustered method (trained and evaluated on MedSpEN)

	P@1	P@5	P@10
Supervised - No clustering	59.29	70.22	73.18
Supervised - 7 clusters	63.9	71.51	73.97

4.4.4 Neural Network Results

The linear model converged to a similar result as the supervised MBC-ICP.
Most of the non-linear models converged to a similar result as the linear model, accept the model with ReLU activation, that converged to a wrong solution.
The results can be seen on Table 7.

Note that we explored a deeper architecture for the non-linear model, as well as different initialization for the models (included MBC-ICP output as the initialization). No effect were observed, and the results were similar.

Table 7: Accuracy (%) comparison between neural network setups (evaluated on Spanish for Medical)

Activation	Loss	P@1	P@5	P@10
No activation	0.039	63.23	77.48	80.58
ReLU	0.065	8.67	21.30	28.25
Leaky ReLU	0.041	63.69	78.00	81.27
Sigmoid	0.055	55.19	72.25	77.08
Tanh	0.040	60.53	75.47	79.95

4.4.5 Convergence of the original MBC-ICP method Examination

When we run the original MBC-ICP on the unsupervised professional terms samples, we noticed that the model is not consistently converge. 3 out of 12 runs on the $5,000$ frequent professional terms failed to converge. When we increased the number of samples, the convergence back to stability. This finding emphasis the contribution of the initialization by the lexicon that was inferred by the definitions similarity, as the model managed to converges on every train with it.

5 Conclusions

In this research we examined two approaches for handling the challenge of translating fine-grained professional terms: projection to the global language space and using a known global translation to learn a specialized translation function.

Regarding the projection to the global language space, we conclude that encoded definitions embeddings are not an accurate representation for the professional terms. Though, those lexicons can be utilize as a stable and qualitative initialization for other methods.
Improving the data processing can be considered in order to improve the performance, for example - use the first sentences of the definition may be more concentrated definition.
Nevertheless, we cannot reject the projection to the global language as an approach for dealing with professional terms translation. Our implementation for this approach achieved low accuracy, but this projection can be learned or implemented in a different way, which may yield the required signal.

Regarding the MBC-ICP variations, we conclude that the regression approach reaches its limit. Both the linear and the non-linear models converged to a similar results, even on supervised models which the training and the testing applied on the same data set. Therefore, we would suggest to consider a canonical methods, which map the embeddings in both languages to a new joint space.
Another conclusion, is that the training data should contain also global and non-professional terms, in order to generalize better and learn the languages morphology.

References

M. Artetxe, G. Labaka, and E. Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 451–462, 2017. doi: 10.18653/v1/P17-1042. URL `https://doi.org/10.18653/v1/P17-1042`.

M. Artetxe, G. Labaka, and E. Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, 2018.

P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, 1992. doi: 10.1109/34.121791. URL `https://doi.org/10.1109/34.121791`.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016. URL `http://arxiv.org/abs/1607.04606`.

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

C. España-Bonet, Á. C. Varga, A. Barrón-Cedeño, and J. van Genabith. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *J. Sel. Topics Signal Processing*, 11(8):1340–1350, 2017. doi: 10.1109/JSTSP.2017.2764273. URL https://doi.org/10.1109/JSTSP.2017.2764273.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.

K. M. Hermann and P. Blunsom. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 58–68, 2014. URL https://www.aclweb.org/anthology/P14-1006/.

Y. Hoshen and L. Wolf. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 469–478, 2018. URL https://aclanthology.info/papers/D18-1043/d18-1043.

Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751, 2014. URL http://aclweb.org/anthology/D/D14/D14-1181.pdf.

G. Lample and A. Conneau. Crosslingual language model pretraining. *arXiv preprint arXiv:1710.04087*, 2019.

N. Limsopatham and N. Collier. Adapting phrase-based machine translation to normalise medical terms in social media messages. *EMNLP*, page 1675–1680, 2015.

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013. URL http://arxiv.org/abs/1309.4168.

N. Nakashole and R. Flauger. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 221–227, 2018. doi: 10.18653/v1/P18-2036. URL https://www.aclweb.org/anthology/P18-2036/.

S. Ruder, I. Vulic, and A. Søgaard. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.*, 65: 569–631, 2019. doi: 10.1613/jair.1.11640. URL https://doi.org/10.1613/jair.1.11640.

H. Schwenk and M. Douze. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 157–167, 2017. URL https://aclanthology.info/papers/W17-2619/w17-2619.

M. Villegas, A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, and M. Krallinger. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. *Language Resources and Evaluation*, 05 2018.

W. Weng and P. Szolovits. Mapping unparalleled clinical professional and consumer languages with embedding alignment. *CoRR*, abs/1806.09542, 2018. URL http://arxiv.org/abs/1806.09542.

W. Weng, Y. Chung, and P. Szolovits. Unsupervised clinical language translation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019.*, pages 3121–3131, 2019. doi: 10.1145/3292500.3330710. URL https://doi.org/10.1145/3292500.3330710.

Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Ábrego, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. Multilingual universal sentence encoder for semantic retrieval. *CoRR*, abs/1907.04307, 2019. URL http://arxiv.org/abs/1907.04307.