*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4157–4166
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

# Multilingualization of Medical Terminology: Semantic and Structural Embedding Approaches

**Long-Huei Chen, Kyo Kageura**
The University of Tokyo
Tokyo, Japan
longhuei@g.ecc.u-tokyo.ac.jp, kyo@p.u-tokyo.ac.jp

### Abstract

The multilingualization of terminology is an essential step in the translation pipeline, to ensure the correct transfer of domain-specific concepts. Many institutions and language service providers construct and maintain multilingual terminologies, which constitute important assets. However, the curation of such multilingual resources requires significant human effort; though automatic multilingual term extraction methods have been proposed so far, they are of limited success as term translation cannot be satisfied by simply conveying meaning, but requires the terminologists and domain experts' knowledge to fit the term within the existing terminology. Here we propose a method to encode the structural properties of terms by aligning their embeddings using graph convolutional networks trained from separate languages. The results show that the structural information can augment the standard bilingual lexicon induction methods, and that taking into account the structural nature of terminologies allows our method to produce better results.

**Keywords:** Terminology, Multilinguality, Machine Learning, MultiWord Expressions, Translation Resources

## 1. Introduction

Terminology occurs along the line between vocabularies and knowledge systems. While many regard terminologies as specialized dictionaries that contain domain-specific terms, terminology exhibits characteristics one does not typically expect from a list of words. Specifically, we consider the prescriptive aspects of terminologies one cannot easily observe in the descriptive process of dictionary creation (Felber, 1984; Kockaert and Steurs, 2015). As terminologies in any language are the depiction of the underlying knowledge system, the designation of a term must reflect the concept it represents in a consistent and accepted manner to facilitate expert communications (Sager, 1990).

While many terminology processing techniques conflate text processing with terminology processing (Pazienza et al., 2005; Chung, 2003; Peñas et al., 2001), we recognize that the unique nature of terminologies provides essential information that can contribute to more effective modeling. We draw inspiration from the work of human terminologists, as they consult both the text and the terminology when conducting actions such as term identification or multilingualization. Domain experts and terminologists do not simply create terms based on their occurrence in the text, but understand that its designation must reflect its position within the domain- and language-specific terminology. This is related to the theoretical understanding that, linguistically, the formation of term candidates is not syntactic but lexical, and that term candidates turn into terms with their incorporation into the terminological system through a conscious, social process (Kockaert and Steurs, 2015).

We therefore hypothesize that domain-specific terminological information can improve the term multilingualization process. Specifically, terminologies exhibit a structural nature not typically expected in general vocabulary (Sager, 1990). We propose a method towards multilingualizing terminologies by taking into account the specialized conceptual structure reflected in existing terminologies. This also amounts to taking into account information on the conceptual system of the domain as opposed to information on domain-specific discourse which has been explored in most term processing (Aker et al., 2007; Morin et al., 2007).

Here we outline the major sections of this paper, which also corresponds to the major contributions.

- Make the case for the need for terminology multilingualization in the industry (§2) and outlines the construction of terminology language resources by using structural information that is part of the terminology (§2.2).
- Measure the limits of semantic-based methods when applied to terminologies (§4).
- Propose a compositional approach to the structural encoding of terminologies, and showcase how multilingualization can assist the inclusion of new terms (§5).
- Harness the full power of our proposed approaches by combining the semantic and structural information for the task of multilingualization (§7).

## 2. Multilingualization of Terminologies

Human-curated, high-fidelity terminology resources are one of the key assets of any translation agency, as they are necessary to ensure the correct translation of specific domain terms among a group of translators. Terminology differs from the general vocabulary in that a translation that conveys equivalent meaning is not necessarily enough, and a set of standardized, high-quality terms is essential for best practices (Gornostay, 2010).

Due to the human curation nature of terminologies, constantly updating them across language barriers requires significant effort (Wright and Budin, 2001). Our work to automate the process can save human resources, as when new terms are extracted from separate languages, translations only need to be verified and not individually translated.

### 2.1. Task Definition

We define the task of terminology multilingualization as the matching between terms in separate languages, which in conjunction produces a multilingual term bank resource for translators and other actors. As terms in a language are created within the existing structure of terminologies of that particular language, making structural correspondences across terminologies in different languages can ensure that proper term candidates are created with little human input.

Publicly, efforts to create, organize, and standardize large-scale terminologies include works of the World Intellectual Property Organization (Valentini et al., 2016), the EU's Inter-Agency Terminology Exchange (Johnson and Macphail, 2000), and more recently the Terminology Working Group of the Japan Translators Federation. We see that our work is not simply translation for translation's stake, but rather an effort to solve a problem faced by translation agencies every day.

### 2.2. The Structural Nature of Terminologies

To further outline the specific nature of terminologies, we compare terminology with other, related, data resources.

- **Terminology and multi-word expressions (MWEs).** A common misconception by many outside (or even inside) the terminology community is the conflation of the two by treating processing of terminologies (of which approximately 80% is complex (Kageura, 2012)) as simply processing of multi-word expressions. But whereas MWEs are characterized by the fact that their meaning cannot be constructed from the meaning of their parts compositionaly, for terminology the constituent elements come together to produce the advanced concept that is highly related to the constituents.
- **Terminology and knowledge graphs:** An often ignored fact is that each term is motivated by an underlying concept of the knowledge domain. This reflects a shared constructive nature between terminologies and knowledge graphs. We can further draw inspiration from the knowledge graph technique stemming from its graph nature, based on the understanding that terminologies in different languages are driven by the same underlying concepts.
- **Multilingual terminology and dictionaries:** While in the creation of dictionaries, the lexicographer's job is to descriptively note the corresponding words or expressions in the respective languages, in creating multilingual terminologies the terminologist must consider both the meaning and the existing domain terminology to ensure the prescribed term fits with the domain knowledge.

### 2.3. Semantic and Structural Information in Terminology Multilingualization

As pointed out in §1, we aim to draw inspiration from the work of human terminologists in the incorporation of terminology structure in the task of multilingualization. We therefore conducted multilingualization experiments using our proposed method of compositional structural embeddings (§5), and compared with bilingual lexicon induction methods where only semantic and no structural information is applied in prediction (§4). In the following section, we describe the overall experiment setup common to the three approaches.

## 3. Experiment Setup

To explore the validity of various approaches in the term multilingualization task as detailed in §2, we describe the data source and the shared experimental setup with details in the following.

### 3.1. Data Source: Medical Subject Headings (MeSH)

| Language Pair | | Term Pairs |
|---|---|---:|
| English (en) | French (fr) | 23461 |
| English (en) | Spanish (es) | 24182 |
| English (en) | German (de) | 24102 |
| English (en) | Russian (ru) | 25847 |
| English (en) | Finnish(fi) | 16425 |
| English (en) | Czech (cs) | 12086 |

Table 1: Language pairs extracted from the metathesaurus provided as part of the UMLS Terminology Services. We selected the language pairs to be used in the experiments by taking into account diverging language ancestry and data sizes.

**Medical Subject Headings (MeSH)** (Lipscomb, 2000) is a medical terminology resource published by the United States National Library of Medicine (NLM). The controlled vocabulary is created mainly to assist in the classification of medical research into a systematic format that allows the retrieval of information accordingly. Translation of the MeSH terminology is available in 10 languages as part of the Unified Medical Language System (UMLS) (Bodenreider, 2004); we selected 6 language pairs, with diverging nature and data sizes as shown in Table **??**, each with term pairs of corresponding translation.

Due to its purpose, MeSH terms are organized in a tree structure that shows the broader to more specific concept relations of medical terms (See Figure 1 for a tiny part of the tree). This structural nature is something we aim to take advantage of in our task of multilingualization.

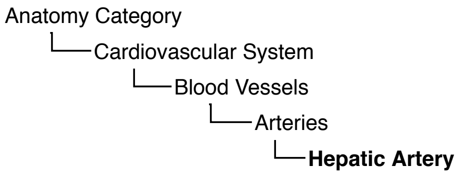


Figure 1: A clip of the tree structure across the terminologies in Medical Subject Headings (MeSH), with `Hepatic Artery` as the example term.

### 3.2. Evaluation: Precision @ *k*

The results are evaluated with precision @ $k$; specifically, for a source term, a match is found when one of the top-$k$ target term nearest neighbor candidates with the smallest distance to the source term is the ground truth. The evaluation scheme is created by keeping in mind the real-world use case where a list of $k$ candidates is provided to the terminologist to assist in the standardization of correct translation. For all language pairs, 10% of term pairs are held out as the test set, and we report all results in the following sections on this test set.

## 4. Semantic Embedding Model: Bilingual Lexicon Induction (BLI)

In the following experiments based on the state-of-the-art model for general domain word translation, we take into account only the semantic information contained in the terms, and apply a bilingual lexicon induction method in a supervised manner to see how well it performs without taking advantage of the structural nature as we proposed in §2.2.

### 4.1. Related Work: Bilingual Lexicon Induction

Bilingual lexicon induction (BLI) approaches can be divided based on the task into supervised (Mikolov et al., 2013; Faruqui and Dyer, 2014; Mikolov et al., 2013) and unsupervised (Barone, 2016), and method-wise to statistics (Artetxe et al., 2018; Artetxe et al., 2017) and adversarial training approaches (Conneau et al., 2017; Patra et al., 2019).

### 4.2. Method: MUSE (Multilingual Unsupervised or Supervised word Embeddings)

Multilingual Unsupervised or Supervised word Embeddings (MUSE) (Conneau et al., 2017) is one of the state-of-the-art methods for bilingual lexicon induction. We selected the method as it has a high degree of support with an open-source library. Despite a few recent works with minor improvements to the model (Patra et al., 2019; Jawanpuria et al., 2019), we decide to apply MUSE due to the fact that it is widely accepted and its performance is ranked at or close to the top in benchmark studies.

The MUSE model takes as input two sets of embeddings, one in each language. The model aims to produce a linear mapping $W$ that can transform a set of embeddings in one language to the other language, such that corresponding translation of words have their embeddings close together in the vector space.

*** Here you need to articulate the sentences again *** While we refer the reader to the full paper (Conneau et al., 2017) for a detailed description and its performances on general vocabulary translation, the key understanding is that we train two neural networks simultaneously, one discriminator that attempts to determine whether an embedding comes from one language ($y_i$) or is transformed from the other ($Wx_i$); and another mapping objective trains a network such that the transformed embeddings can fool the discriminator as the transformed embeddings $Wx_i$ of one language is close in space to the $y_i$ of the other language's. The discriminator is defined as:

$$\begin{aligned}\mathcal{L}_D\left(\theta_D \mid W\right) = & -\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_D}\left(\text { source }=1 \mid W x_i\right) \\ & -\frac{1}{m} \sum_{i=1}^{m} \log P_{\theta_D}\left(\text { source }=0 \mid y_i\right)\end{aligned} \tag{1}$$

where $\theta_D$ is the discriminator parameters. The mapping obdjective is:

$$\begin{aligned}\mathcal{L}_W\left(W \mid \theta_D\right) = & -\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_D}\left(\text { source }=0 \mid W x_i\right) \\ & -\frac{1}{m} \sum_{i=1}^{m} \log P_{\theta_D}\left(\text { source }=1 \mid y_i\right)\end{aligned} \tag{2}$$

#### 4.2.1. Pre-trained Embeddings from Crosslingual Language Models

While the original task of unsupervised bilingual lexicon induction refrains from techniques that are built on parallel corporal resources, in our terminology applications we aim to exploit the power of cross-lingual pre-trained language models. To do so we extracted pre-trained feature embeddings for the terms from the following 3 multilingual models, multilingual **BERT** (Devlin et al., 2018), XLM-RoBERTa (**XLM-R**), and **FastText** (Bojanowski et al., 2017; Joulin et al., 2016).

- **Multilingual BERT** (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is simultaneously pre-trained on multilingual corpora with a shared vocabulary across languages. The model is pre-trained on a transformer model with objectives on masked language model (MLM) tasks based on a very large-scale general corpus.
- **XLM-RoBERTa (XLM-R)**, or Cross-lingual Language Model pretraining–Robustly Optimized BERT pretraining Approach (Liu et al., 2019), extends the RoBERTa model, which itself is a more advanced evolution of the BERT model, to multilingual training objectives [1].
- **FastText** (Bojanowski et al., 2017; Joulin et al., 2016) pre-trained word vectors are the top of the crop of the previously dominant word-based vectors. Unlike the previous two, these embeddings are not context-sensitive, and also do not consist of tokens and therefore are based on words.

Terms are produced from $n$-gram tokens for both the Multilingual-BERT and XLM-RoBERTa (XLM-R) models, where the embedding for a term is extracted from the second-to-last layer of the model and mean-pooled across

[1] At the time of writing, the full cross-lingual pre-trained model had not been released publicly. We still aimed to compare the improved crosslingual model features with the multilingual BERT but acknowledge that the results here are not based on the full model and are subject to changes in the near future.

individual tokens (since the tokens are character *n*-grams for both models, single- and multi-worded terms are not distinguished here). For FastText, we similarly mean-pool the individual word embeddings for terms.

### 4.3. Results and Discussion

We set up the experiments as outlined in §3. We applied the state-of-the-art MUSE method on top of three different types of term embeddings created on pre-trained models detailed in §4.2.1. The results are shown in Table 2 (a).

We found that the performance is better across all embedding sources for language pairs with a traditionally larger amount of parallel language resources. Despite the more limited scope of the *n*-grams vocabulary for both BERT and XLM-R, they generally outperformed the word-based fastText embeddings, suggesting the validity of the *n*-gram approach to token embedding generation.

Finally, we see that FastText embeddings, despite the far larger pre-trained vocabulary, do not generally produce better results due to the limited coverage among medical terms (which are often rare words); this is in comparison with the *n*-gram based language models where all terms can be covered with a pooled embedding. For more discussion, we refer the reader to §5.4, where we compare language pairs across both semantic and structural embedding methods.

## 5. Structural Embedding Model: Graph Convolutional Networks

### 5.1. Related Work

Bilingual term extraction from parallel and comparable corpora has been studied since the early 1990s (Dagan and Church, 1994; Daille et al., 1994; Morin et al., 2007). Some use compositional translation to match bilingual terms (Tonoike et al., 2005; Delpech et al., 2012). More recently, word embeddings are used for bilingual terminology extraction (Hazem and Morin, 2017).

The terminological network is defined as a network where each node is a term (single- and multi-worded) and two nodes share an edge when they have one or more constituent elements in common (Iwai et al., 2016b). Such networks have been proven useful in subdomain delineation, and the generation of multi-word terminologies (Iwai et al., 2016a).

Graph convolutional networks (GCNs) (Kipf and Welling, 2016) process a graph by generating a feature representation of each node. They enable the numerical encoding of terminological graph structure into a high-dimensional space. They have been applied to many real-world data that is naturally represented as graphs, and have found use in text classification (Schlichtkrull et al., 2018), protein interface prediction (Fout et al., 2017), and semantic role labeling (Marcheggiani and Titov, 2017).

For knowledge graphs, alignments between graphs have been an active topic for research. The advent of embeddings methods (Wang et al., 2014) brought new life to the community, with entity embeddings created from attributes and graph structures such as MTransE (Chen and Zaniolo, 2017), TransR (Huang et al., 2017), and TransD (Ji et al., 2015). More recently, graph convolutional networks have been applied to knowledge graphs and extended to cross-lingual settings (Chen et al., 2016; Shang et al., 2019; Xu et al., 2019).

### 5.2. Method: Cross-lingual Graph Alignment via Graph Convolutional Networks

We begin by generating the initial structural embedding of terms by tokenizing and pooling the token embeddings, which are trained from scratch. We then train two graph convolutional networks on top of the term embeddings separately for each language according to the terminology tree, optimizing to minimize the output feature distances and thus aligning the graphs. Finally, we test out the task of multilingualization by predicting the translation of terms unseen during training.

### 5.3. Tokenization of Terminologies to Generate Structural Embeddings

The key advancement we make is the creation of *n*-gram-based token-individual embedding in each language that, when combined to form a term embedding, is trained to reflect the term's place in the terminology tree structure (§3.1). This follows the framework:

1. Tokenize the single- and multi-word terms into character *n*-gram tokens.

2. For each token, assign a randomly initialized embedding.

3. The pooling of token embeddings produces the embedding of the term, which is trained to reflect its position in the terminology tree structure.

4. For a new term in one language, the sum of the token embedding represents its position in the terminological structure, and can be used to find the corresponding term in the other language's aligned graph.

We applied the same tokenizer from the BERT model to make the results comparable. For each token, we used the defined adjacency matrix along with the graph convolutional network to align the token embeddings with the overall graph.

#### 5.3.1. Graph Convolutional Network (GCNs) Building Blocks

The basic building block of the graph convolution operation is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

where $H$ is the node features from the previous layer, $A$ is the adjacency matrix, $W^{(l)}$ is the trainable weight for the current convolutional layer $(l)$, and $\sigma$ is an nonlinear activation function, in this case $\mathrm{ReLU}$. $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix, where $\tilde{D}$ is the diagonal node degree matrix.

Typically for each layer, the weights convolve the node features to a smaller dimension, allowing for an effective convolutional operation that produces higher-level features. This allows the eventual node features to encode the neighborhood and graph topology information.

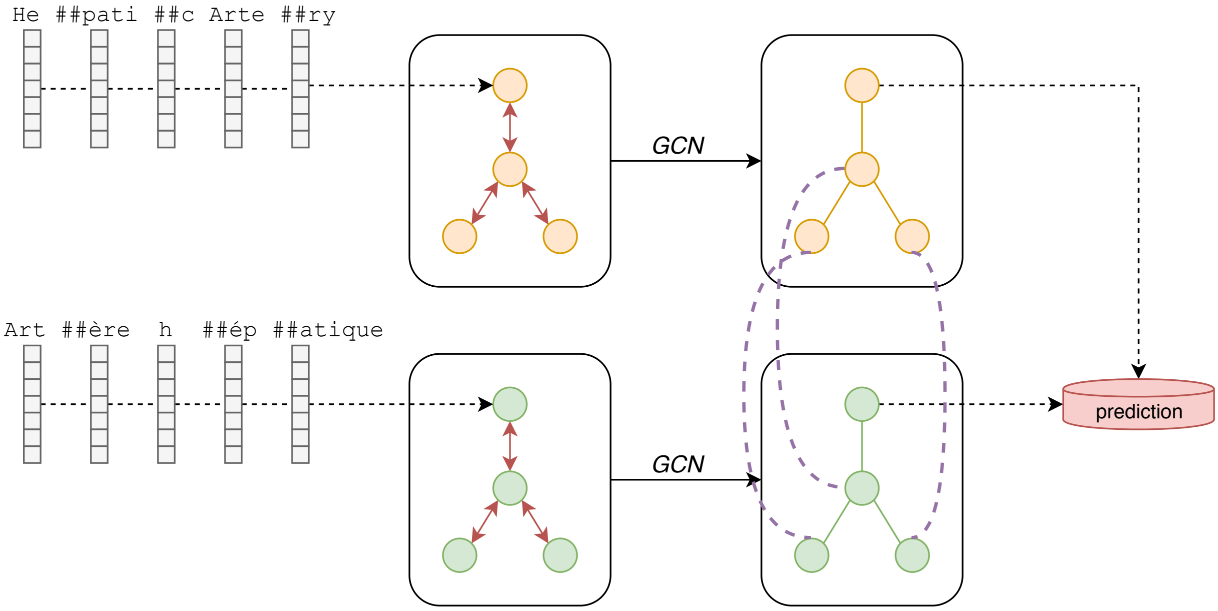


Figure 2: The model architecture (§5) where token embeddings are trained with a two-layer graph convolutional network (GCN) such that when combined, they reflect the underlying graph structure of the terminology. As the two GCN individual to the languages are aligned, the approach allows prediction of previously unseen terms to fit within the terminology tree structure, and thus multilingualization based on their position.

### 5.3.2. GCNs for Terminology Tree

For our task of multilingualization, we trained two 2-layer GCNs, one for each language, such that the embeddings for the terms reflect the tree structure. Figure 2 shows a schematic of our approach. The two GCNs are aligned with one another in a way that corresponding term pairs come close in space, so when a new term is generated (also from the pooling of trained embeddings) it can find corresponding terms in the other language.

To simplify the model training, the tree of the terminology base is converted into graph adjacencies such that each term is connected to both its immediate parent and children in the tree, thus the token embedding reflects basically the position and ancestral information of the term concept in the tree structure.

### 5.3.3. Loss Function and Prediction

The loss function is the distance function calculated between corresponding terms in both languages (the positive samples), subtracted by the distances between the negative samples, which are randomly selected terms from both languages that are not translations of each other.

$$L = \sum_{(s,t)\in S,(s',t')\in S'} [\ \|\boldsymbol{e}(s) - \boldsymbol{e}(t)\|_1 \\ - \|\boldsymbol{e}(s') - \boldsymbol{e}(t')\|_1]$$

where $S$ is the set of positive samples, $S'$ the negative set, and $\boldsymbol{e}(i)$ is the embedding vector of the term $i$. Prediction is similarly carried out across all terms of respective languages with the distance function

$$D(s,t) = \|\boldsymbol{e}(s) - \boldsymbol{e}(t)\|_1$$

and we report $k = n$ results with the smallest distance and thus the highest similarities.

The model is optimized with an Adam optimizer (Kingma and Ba, 2014) with learning rate $= 1$. The token embeddings are randomly initialized with dimension 300 and passed through two graph convolutional layers with hidden dimension 200, and towards an output layer with dimension set to 100.

## 5.4. Results and Discussion

We compared results for each language pair in both directions across all the embeddings for the semantic methods and our proposed structural embedding method (Table 2 (b)). The best results for each language pair are indicated in bold.

We found that the performance is better across all embedding sources for language pairs with a traditionally larger amount of parallel language resources. Despite the more limited scope of the $n$-grams vocabulary for both BERT and XLM-R, they generally outperformed the word-based FastText embeddings, suggesting the validity of the $n$-gram approach to token embedding generation.

We can see that our structural embedding-based model almost always outperforms the semantic method for $k = 10$ cases. This is a set up that mirrors the working reality of terminologists, as a list of target term candidates can be presented for human decision-making in multilingual terminology curation. The model suggestions could potentially save many hours of labor in the pinpointing of the correct translation of a term. The semantic methods generally perform better when we limit to $k = 1$ cases, corresponding to fully automated target term candidate generation.

|  | (a) Semantic |  |  |  |  |  | (b) Structural |  |
|---|---|---|---|---|---|---|---|---|
| Embeddings | BERT |  | XLM-R |  | fastText |  | Structural |  |
| Precision @ $k$ | $k = 1$ | $k = 10$ | $k = 1$ | $k = 10$ | $k = 1$ | $k = 10$ | $k = 1$ | $k = 10$ |
| en-fr | **42.14** | 56.46 | 35.19 | 47.72 | 31.27 | 48.06 | 19.55 | **64.25** |
| fr-en | **41.97** | 56.50 | 35.32 | 46.14 | 31.83 | 49.08 | 19.27 | **64.25** |
| en-es | **43.49** | 58.78 | 37.29 | 49.94 | 35.43 | 53.58 | 20.03 | **59.07** |
| es-en | **43.53** | 57.88 | 34.44 | 47.04 | 37.33 | 53.78 | 19.84 | **59.07** |
| en-de | **49.19** | **62.71** | 40.52 | 51.51 | 23.06 | 38.53 | 17.42 | 55.12 |
| de-en | **49.27** | **62.59** | 41.85 | 52.68 | 25.88 | 39.44 | 17.65 | 55.12 |
| en-ru | **16.40** | 26.15 | 14.00 | 21.90 | 7.35 | 13.42 | 15.56 | **51.53** |
| ru-en | **16.56** | 26.73 | 12.92 | 20.15 | 6.03 | 11.33 | 16.50 | **51.53** |
| en-fi | **34.63** | 47.29 | 28.30 | 39.93 | 9.74 | 20.75 | 19.61 | **58.08** |
| fi-en | **33.29** | 45.59 | 29.03 | 38.59 | 9.07 | 19.84 | 19.47 | **58.08** |
| en-cs | **50.21** | 62.45 | 41.27 | 52.11 | 20.26 | 36.97 | 21.60 | **64.52** |
| cs-en | **48.30** | 61.70 | 41.11 | 50.37 | 23.49 | 42.27 | 21.88 | **64.52** |

Table 2: Results based on (a) the bilinglingual lexicon induction (BLI) method, specifically MUSE, as outlined in §4, and (b) structural methods based on graph convolutional networks with compositional encoding of the terminologies (§5). We report the top-$k$ results where the $k$ target terms with highest similarities to the source term are considered. The best results among all methods are indicated in bold.

| Top-$k$ Match | en-fr | fr-en | en-es | es-en | en-de | de-en | en-ru | ru-en | en-fi | fi-en | en-cs | cs-en |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $k = 1$ | 20.93 | 20.12 | 21.04 | 20.30 | 19.58 | 18.98 | 6.40 | 5.46 | 19.68 | 20.01 | 24.72 | 26.19 |
| $k = 10$ | 67.90 | 67.80 | 62.89 | 62.33 | 61.15 | 60.97 | 27.11 | 25.35 | 57.47 | 58.62 | 69.85 | 69.21 |
| $k = 50$ | 92.85 | 93.09 | 90.49 | 90.40 | 90.05 | 90.60 | 59.00 | 55.35 | 91.14 | 90.81 | 85.01 | 84.74 |

Table 3: Results from the fusion method combining the semantic and structural embeddings. Compared with results from individual embeddings in Table 2 we see a general trend of improvement.

# 6. Analysis

## 6.1. Selection of *k*

To further delineate cases where our models can help the terminologists in determining the best translation, we plot the precision against the $k$ candidates that are the nearest neighbors to the source terms. We show three of the language pairs with unique patterns in Figure 3. Most language pairs follow the same pattern as for en-fr (3a), i.e. whereas the bilingual lexion induction (BLI) methods outperform our model when k is smaller than five, our model has the potential to suggest candidates that better encompass the correct translation when $k$ is raised slightly.

For the English/German language pair (Figure 3b), our model surpasses the state-of-the-art only at $k$ around 15 rather than a smaller value like the others. We believe this is due to the unique abundance of compound nouns as terms in German, and since the tokenization engine of our model is based on BERT, the limited quality of the tokenization does not allow our model to fully learn the tree with compositional embeddings. For bilingual lexion induction the embedding for terms is obtained from the average of token embeddings and is less affected by the tokenization composition.

A special case must also be made for lower-resource languages; as the multilingual BERT and XLM-RoBERTa embeddings do not make a distinction based on input text language, its capabilities can differ among languages. In the case of Russian (Figure 3c), where Cyrillic and not Latin script is used, the results are especially poor, showcasing the differential coverage of the multilingual model, an issue that is beginning to attract attention in the community recently (Libovický et al., 2019).

Overall we conclude that the semantic-based BLI methods, when applied to terminology multilingualization, see room for improvement as they do not improve in precision even when $k$ is set to be large. Our method is able to surpass the bilingual lexion induction methods when $k$ is sufficiently large. As such, we can see that our method successfully makes use of the information contained in terminologies. Whether this is due to the incorporation of domain-specific elements or of structural nature is, as of now, not clear. More work is required to improve on the precision by further diagnosing the results and exploring structural embeddings which capture the position of terms in terminologies with greater precision.

## 6.2. Tree Depth and Qualitative Analyses

To clarify the nature of terms where the translation is correctly returned, we ploted the average precision @ 10 for all language pairs against the term's depth in the structural terminology tree, as shown in Figure 4. We can see that pre-

|  |  | (a) Proposed Model (Ours) |  |
|---|---|---|---|
|  |  | Correct | Incorrect |
| (b) Semantic (BERT) | Correct | Abortion Applicants/Femmes demandant l'avortement | Bone Nails/Clous orthopédiques |
|  |  | Sugar Acids/Oses acides | Jejunal Disease/Maladies jéjunales |
|  |  | Food, Organic/Nourriture biologique | Weed Control/Lutte contre les mauvaises herbes |
|  |  | Caspases, Initiator/Caspases initiatrices | Peanut Agglutinin/Lectine cacahuète |
|  |  | Sodium Compounds/Composés du sodium | Ergothioneine/Thionéine |
|  | Incorrect | Emetics/Agents émétiques | Ankle Brachial Index/Index de pression systolique |
|  |  | Mitogen-Activated Protein Kinase 3/MAPkinase-3 | Neurturin/Protéine Nrtn |
|  |  | Paragonimiasis/Distomatose à Paragonimus | Anostraca/Crevettes féeriques |
|  |  | DNA, Helminth/ADN des helminthes | Dental Staff, Hospital/Personnel dentaire hospitalier |
|  |  | Interleukin-8/AMCF-I | Ethnic Cleansing/Épuration ethnique |

Table 4: Matrix of select samples of English/French term pairs that are both or either correctly and incorrectly translated with (a) our structural embedding model, and (b) the BLI-based methods. We observe that the semantic embedding model tend to translate commonly-used words and phrases correctly, while for our structure-based method no such distinction is found.

cision is higher for our method up to level 3, which contains more general concepts, whereas the BLI methods based on BERT embeddings catch up for terms nearer the more specified terms at or nearer the leave nodes.

We postulate that our structural embeddings, as they are based only on structural information, perform better in cases of capturing the nature of terms that represent more general concepts in a more-fine grained way, which BLI methods fail to do. On the other hand, the semantic method can take better advantage of a wider range of information for decision-making at a more specific concept level. The diverging strength of our model compared to the bilingual lexicon induction method suggests that a model that combined both semantic and structural information could push performances further.

In Table 4 we list example term pairs in the English-to-French languages that are either correctly translated by (a) our proposed model or (b) the MUSE method based on BERT embeddings, or both, or neither. For the semantic methods, we can see that term pairs where the model performs correctly are cases where the terms are common phrases that are not limited to the medical domain, while expert terms are more difficult for the model to get right, reflecting the method's original design for general word translation. For our model we do not see such a deficiency regarding medical terms.

## 7. Semantic + Structural Fusion Embeddings Model

In the final part of our experiments, we combined the two aforementioned approaches to fuse the semantic and structural terminology information to observe how the joined forces can further enhance terminology multilingualization.

### 7.1. Method

The fusion model creates a prediction based on both the BERT embeddings transformed with the MUSE method (§4) and the structural embedding obtained from our proposed GCN model (§5); specifically, between candidate source term $s$ and target term $t$:

$$\begin{aligned} D(s,t) =& \beta \frac{\|\boldsymbol{e_{sem}}(s) - \boldsymbol{e_{sem}}(t)\|_1}{d_{sem}} \\ &+ (1-\beta) \frac{\|\boldsymbol{e_{stc}}(s) - \boldsymbol{e_{stc}}(t)\|_1}{d_{stc}} \end{aligned}$$

where $\boldsymbol{e_{sem}}$ are the semantic embeddings and $\boldsymbol{e_{stc}}$ are the structural embeddings, $d$ is the respective dimensions of the embeddings, and $\beta$ is a hyperparameter. The top-$k$ matches with the lowest distances between the languages are then selected and accuracies calculated from cases where the ground truth is in one of the matches.

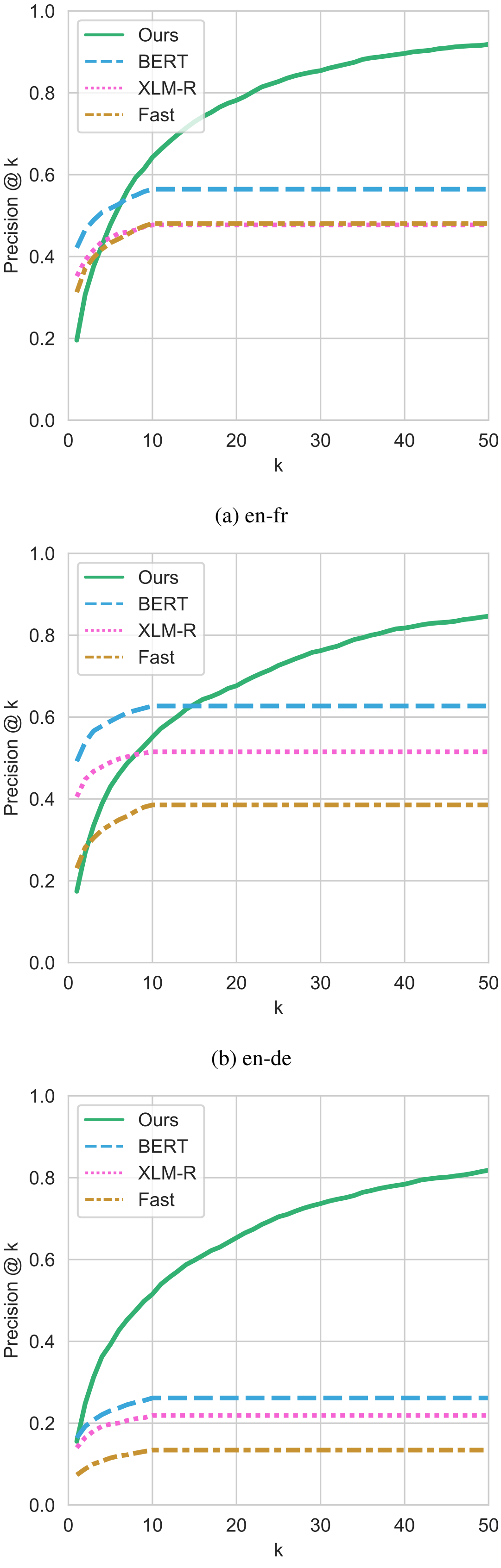
### 7.2. Results & Discussion

We found that in general, the fusion model produces slightly better results compared to the individual embeddings models for $k = 10$ (Table 3). We recognize that this presents a chance to enhance the embeddings by combining the semantic and structural information present in the terminologies. Specifically for $k = 1$ cases, the fusion model underperforms, which is possibly due to the simplistic nature of the approach which results in the fusion process not taking full advantage of the strengths offered by each.

The limited improvements to the results suggest the need for a more advanced fusion model where the semantic and structural models are trained in tandem instead of combining the results after the fact. The efforts to tackle this is our ongoing work, based on our observation that these methods based on semantic and structural information complement each other.

## 8. Conclusion

Terminology processing differs from most language processing in the requirement of preciseness, precision, and consistency, reflecting its theoretical position. Our proposed model serves the needs of terminologists in providing a consistent approach to assist and partially automate multilingualization with a short-list of candidates. We have confirmed the viability of structural information being applied to the task of multilingualization, which is held to be an essential step in any translation process involving terminologies.



(a) en-fr

(b) en-de

(c) en-ru

Figure 3: Precision varying w.r.t. $k$ from 1 to 50 for three representative language pairs. For most language pairs our model initially has lower precision but surpasses the semantic methods quickly at $k$ around 5 to 10, a pattern shown in (a).

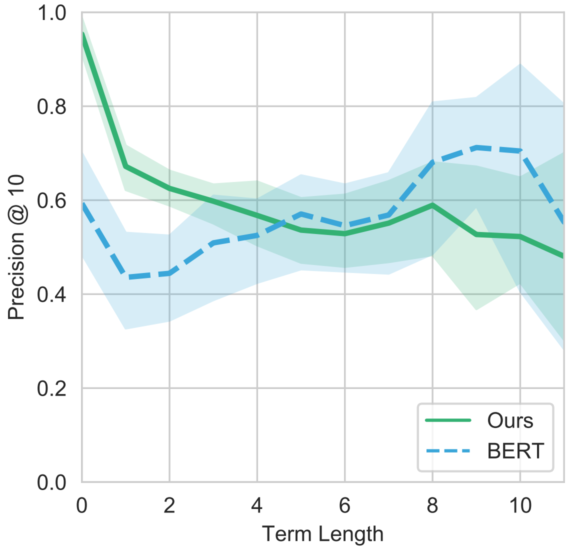


Figure 4: Precision @ 10 with respect to term location in tree for all language pairs. The two models complement each other in strength regarding more general and more precise term concepts.

With the aforementioned challenges in mind, we are currently working to:

- make the performance consistent with respect to precision @ $k$, to fully harness the power of the domain-specific structural information;
- extend our model to allow simultaneous learning based on the fusion of semantic and structural information inherent in the terminology tree; and
- generalize the model to generate terms independent of a list of prescribed candidates, such that new terms can be translated even when the concept does not yet exist in the target language.

Semantic information is applied in our proposed model only in the sense that the tokenziation learned from a general, domain-nonspecific language model is applied to develop the composition. We do see, however, that a domain-specific knowledge model can help produce a fusion model that takes advantage of both the semantic and structural domain-specific information. Moreover, more work is needed to make clear the structural information components that contribute directly to the improved performance within the overall domain-specific training apparatus.

## 9. Acknowledgements

We thank Dr. Kazawa Hideto for his invaluable advice and comments in this ongoing work. This work is partly supported by JSPS KAKENHI 19H05660.

## 10. Bibliographical References

Aker, A., Paramita, M., and Gaizauskas, R. (2007). Extractiing bilingual terminologies from comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 664–671.

Artetxe, M., Labaka, G., and Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Barone, A. V. M. (2016). Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv preprint arXiv:1608.02996*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chen, M. and Zaniolo, C. (2017). Learning multi-faceted knowledge graph embeddings for natural language processing. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 5169–5170.

Chen, M., Tian, Y., Yang, M., and Zaniolo, C. (2016). Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *arXiv preprint arXiv:1611.03954*.

Chung, T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2):221–246.

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

Dagan, I. and Church, K. W. (1994). Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 34–40.

Daille, B., Gaussier, E., and Lange, J. M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *COLING 1994*, pages 515–521.

Delpech, E., Daille, B., Morin, E., and Lamaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In *COLING 2012*, pages 1–17.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Felber, H. (1984). *Terminology manual*. Infoterm.

Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. In *Advances in Neural Information Processing Systems*, pages 6530–6539.

Gornostay, T. (2010). Terminology management in real use. In *Proceedings of the 5th International Conference Applied Linguistics in Science and Education*, pages 25–26.

Hazem, A. and Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 685–693.

Huang, W., Li, G., and Jin, Z. (2017). Improved knowledge base completion by the path-augmented transr model. In *International Conference on Knowledge Science, Engineering and Management*, pages 149–159. Springer.

Iwai, M., Takeuchi, K., and Kageura, K. (2016a). Cross-lingual structural correspondence between terminoogies: The case of english and japanese. In *Proceedings of the 12th International conference on Terminology and Knowledge Engineering (TKE)*, pages 14–23.

Iwai, M., Takeuchi, K., Kageura, K., and Ishibashi, K. (2016b). A method of augmenting bilingual terminology by taking advantage of the conceptual systematicity of terminologies. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 30–40.

Jawanpuria, P., Balgovind, A., Kunchukuttan, A., and Mishra, B. (2019). Learning multilingual word embeddings in latent metric space: a geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

Ji, G., He, S., Xu, L., Liu, K., and Zhao, J. (2015). Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kageura, K. (2012). *The quantitative analysis of the dynamics and structure of terminologies*, volume 15. John Benjamins Publishing.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kockaert, H. J. and Steurs, F. (2015). *Handbook of terminology*, volume 1. John Benjamins Publishing.

Libovickỳ, J., Rosa, R., and Fraser, A. (2019). How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marcheggiani, D. and Titov, I. (2017). Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

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

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 402–411.

Patra, B., Moniz, J. R. A., Garg, S., Gormley, M. R., and Neubig, G. (2019). Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. *arXiv preprint arXiv:1908.06625*.

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

Peñas, A., Verdejo, F., Gonzalo, J., et al. (2001). Corpus-based terminology extraction applied to information access. In *Proceedings of Corpus Linguistics*, volume 2001, pages 458–465. Citeseer.

Sager, J. C. (1990). *Practical course in terminology processing*. John Benjamins Publishing.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Shang, J., Ma, T., Xiao, C., and Sun, J. (2019). Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T., and Sato, S. (2005). Effect of domain-specific corpus in compositional translation estimation for technical terms. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 116–121.

Valentini, C., Westgate, G., and Rouquet, P. (2016). The pct termbase of the world intellectual property organization. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 22(2):171–200.

Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1591–1601.

Wright, S. E. and Budin, G. (2001). *Handbook of terminology management: application-oriented terminology management*, volume 2. John Benjamins Publishing.

Xu, K., Wang, L., Yu, M., Feng, Y., Song, Y., Wang, Z., and Yu, D. (2019). Cross-lingual knowledge graph alignment via graph matching neural network. *arXiv preprint arXiv:1905.11605*.

## 11. Language Resource References

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270.

Johnson, I. and Macphail, A. (2000). Iate–inter-agency terminology exchange: Development of a single central terminology database for the institutions and agencies of the european union. In *Proceedings of the Workshop on Terminology resources and computation, LREC 2000*.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.