# BOLT-K: Bootstrapping Ontology Learning via Transfer of Knowledge

Nikhita Vedula
The Ohio State University
vedula.5@osu.edu

Pranav Maneriker
The Ohio State University
maneriker.1@osu.edu

Srinivasan Parthasarathy
The Ohio State University
srini@cse.ohio-state.edu

## ABSTRACT

Dynamically extracting and representing continually evolving knowledge entities is an essential scaffold for grounded intelligence and decision making. Creating knowledge schemas for newly emerging, unfamiliar, domain-specific ideas or events poses the following challenges: (i) detecting relevant, often previously unknown concepts associated with the new domain; and (ii) learning ontological, semantically accurate relationships among the new concepts, despite having severely limited annotated data. To this end, we propose a novel LSTM-based framework with attentive pooling, BOLT-K, to learn an ontology for a target subject or domain. We bootstrap our ontology learning approach by adapting and transferring knowledge from an existing, functionally related source domain. We also augment the inadequate labeled data available for the target domain with various strategies to minimize human expertise during model development and training. BOLT-K first employs semantic and graphical features to recognize the entity or concept pairs likely to be related to each other, and filters out spurious concept combinations. It is then jointly trained on knowledge from the target and source domains to learn relationships among the target concepts. The target concepts and their corresponding relationships are subsequently used to construct an ontology. We extensively evaluate our framework on several, real-world bio-medical and commercial product domain ontologies. We obtain significant improvements of 5-25% F1-score points over state-of-the-art baselines. We also examine the potential of BOLT-K in detecting the presence of novel kinds of relationships that were unseen during training.

**ACM Reference Format:**
Nikhita Vedula, Pranav Maneriker, and Srinivasan Parthasarathy. 2019. BOLT-K: Bootstrapping Ontology Learning via Transfer of Knowledge. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313511

## 1 INTRODUCTION

Ontologies, taxonomies or knowledge graphs represent an effective way of organizing massive amounts of real world information into a structured format. In particular, domain-specific ontologies are valuable resources that formally model the conceptual vocabulary of a given domain. Building accurate ontologies from trustworthy sources [36, 51, 65, 66, 69] with sufficient coverage of concepts and relationships among them is time-consuming and labor-intensive. A number of approaches to automate this process have been proposed. For instance, building ontologies based on statistical, linguistic and graphical features [45, 68, 76]; enriching existing ontologies with domain-specific information [63]; and embedding-based methods to complete knowledge graphs [2, 23, 31, 47, 61, 62, 70, 73, 79, 80]. Nevertheless, they only utilize textual corpora and other accompanying information from the specific domain under consideration. They do not leverage the abundant, hierarchically structured knowledge that might be available in functionally and/or semantically *similar* or *related* subjects or domains. This is especially valuable in fields ranging from epidemiology to crisis response, and from bio-medicine to commercial product catalogs, where emerging knowledge can be frequent and crucial. For instance, there is no semantically coherent ontology associated with the recently surfaced human disease of *Zika fever*, knowledge of which is evolving to-date. It will therefore be highly useful to take advantage of its connections to similar vector-borne diseases like *Dengue* or *Malaria*, for which well organized and annotated information from domain experts is available. Further, there are multiple challenges associated with building comprehensive ontologies for e-commerce product platforms [11, 29]. Numerous closely related product catalogs often have incomplete or incorrectly labeled attributes. Using annotations from semantically similar product listings (such as products from the same category) can help embellish existing listings with missing attributes, as well as detect errors in the labeled attributes. In our work, we *bootstrap* the task of learning ontologies for novel domains or subjects, by adapting and *transferring* existing knowledge from related domain ontologies. This also alleviates the requirement of human expertise to obtain sufficient labeled data on the newly emergent subject.

A crucial task in constructing ontologies is learning hierarchical relationships among multiple concepts from unstructured text. This task requires large amounts of annotated training data associated with ontological concepts and their corresponding relationships. This process is time-consuming, expensive, and necessitates a significant amount of expert knowledge to categorize associations in niche domains that a layperson is unlikely to know about. To resolve this issue, Mintz et al [42] heuristically aligned texts with knowledge graphs via distant supervision to automatically generate training examples. Lin et al [32] applied distant supervision with sentence-level selective attention, while Zeng et al [79] coupled it with multi-instance learning to learn relationships. However such efforts either do not address the problem of the highly imbalanced occurrence of different relationships [53], or require a sizable number of sentences connecting related entities. Another popular strategy to address the issue of insufficient labeled data is *data augmentation*. This technique artificially expands labeled training


*WWW '19, May 13–17, 2019, San Francisco, CA, USA*

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313511

sets by generating new data points or transforming the existing ones, such that the class label properties are preserved [22, 35, 52].

In this paper, we propose a framework *BOLT-K* – Bootstrapping Ontology Learning via Transfer of Knowledge. It uses a long short term memory (LSTM) neural network with attentive pooling, to learn an ontology hierarchy for a given *target* domain. We first obtain all the concepts that constitute the target ontology. We subsequently devise an approach based on semantic and topological attributes to identify the concept pairs likely to be connected by a relationship, and eliminate the remaining spurious combinations. To address the issue of limited relationship-labeled training data, we utilize publicly available textual corpora and ontological information from a functionally similar *source* domain. We also employ data augmentation techniques to generate additional training examples for the target domain. We train our model jointly on the target and source hierarchy information, by sharing the hidden feature representations and appropriate model parameters among them. Finally, we predict ontological relationships between the concepts of the target domain, to construct an ontology for it. We extensively evaluate our framework on several real-world datasets, highlighting the transferability of concepts across comparable subjects. We show that BOLT-K can significantly improve the quality of learned ontologies over state-of-the-art baselines, when bootstrapped with relevant knowledge from a similar subject or domain.

To summarize, the key contributions of our work are:

- We develop a flexible and generalizable framework, BOLT-K, to automatically learn ontologies for contemporary novel or emergent sub-domains of rapidly evolving fields such as bio-medicine, epidemiology, e-commerce and crisis response.
- We propose to transfer existing knowledge from functionally similar domains, and augment the insufficient labeled target training data. This significantly lessens the need for manual expertise during model development and training.
- We extensively evaluate BOLT-K on real-world datasets for various sub-domains within bio-medicine and product graphs. We also show BOLT-K's capability in detecting novel types of relationships that were unseen during training.

## 2 RELATED WORK

The following lines of research are related to our work: (i) augmentation of textual training data; (ii) learning ontological relationships; and (iii) transfer learning for natural language processing tasks.

**Training data augmentation**: This is a common and successful technique in the image processing literature [12, 52, 55], and is slowly gaining popularity in NLP applications [22, 35, 52]. A caveat though, is that supplementing text data via transformations, interpolations or affine perturbations is not as straightforward as performing these enhancements for images. Kafle et al [24] proposed two methods of data augmentation for visual question answering: (i) using semantic segmentation annotations with labels to synthesize certain kinds of questions; and (ii) training a stacked LSTM model to generate questions about images. A common technique of augmenting text is to replace words or phrases with their synonyms from a thesaurus [82], or with appropriate n-grams from a language model [52]. Another technique is to translate sentences into a second language, and then translate them back into the original language to obtain a slight variation of the original sentence [13, 54, 71]. Progress has also been made in developing generative models based on variational autoencoders for sentence generation [7, 21, 58]. In this work, we employ a combination of text substitutions for data augmentation, as specified in Section 4.2.

**Relation Extraction**: Deep learning models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and graph-RNNs have been successful in automatically learning features for extracting semantic relations between a pair of entities [49, 57, 72, 79]. A number of approaches have also employed attention mechanisms in conjunction with RNNs for relation extraction and classification [32, 84]. Another way of solving this problem is by formulating it as a link prediction problem in knowledge graphs [2, 23, 31, 47, 61, 70, 73]. Since our method addresses relation extraction at the level of a sentence or a group of sentences, we find that a bidirectional LSTM with attention-based pooling works well for our purpose. Similar to our work, there have been efforts to extract relations with an insufficient number of labeled examples per relationship type. Yuan et al [78] formulated this as a one-shot classification problem and solve it using a convolutional siamese neural network. Levy et al [30] turned it into a reading comprehension problem by associating multiple textual questions with each relation type and learning answers for them.

**Transfer learning**: Zhang et al [81] have surveyed various cross-dataset transfer learning techniques. This includes the kind of knowledge transfer paradigm that our work addresses, i.e. minimizing the generalization error in the target domain with the help of training instances from two disjoint source and target domains. Transfer learning has been highly useful in low-resource domains such as bio-medicine [26, 40]. Ganin et al [17] introduced a representation learning approach for domain adaptation by adding a gradient reversal layer in a feed forward neural network. They trained their model on a document sentiment classification task using labeled data from a source domain and unlabeled data from a target domain. Long et al [34] presented another approach that learns an unsupervised residual function to adapt classifiers from a source domain to a target domain. Similar to our work, prior efforts addressed the sharing of structural parameters across multiple task domains [1]. Yang et al [74] developed a framework for three different types of transfer learning paradigms in hierarchical RNNs for a sequence tagging task; namely cross-domain, cross-application and cross-lingual transfer. Knowledge transfer has also been used to inform inter-related NLP tasks such as named entity recognition, part-of-speech tagging, chunking and word segmentation [9, 48].

## 3 PROBLEM FORMULATION

We propose to learn an ontology for a *target* domain for which labeled information is very limited, by transferring potentially useful semantic and ontological knowledge from a distinct but related *source* domain. We refer to each element of the ontologies associated with these domains as an *entity* or a *concept*.

The inputs to our problem are:

(1) A knowledge ontology $S$ consisting of entities, categories and labeled relationships $S_R$ among them, for a source domain

(2) A small number of concept pairs that will be part of the ontology $T$ for the target domain. These have been labeled with their respective relationship type from a set of target relationship types $T_R$. Since the target is an emergent domain with little to no labeled information available, we only require at least one labeled concept pair per relation type in $T_R$. Note that though $S$ and $T$ are from related domains, there may or may not be relationship types common to both $S$ and $T$. This is not a requirement for our method.

(3) Corpora of text documents whose sentences contain occurrences of the concepts in $S$ and concepts associated with the target domain, which are to be part of $T$. These documents can either be expert-authored such as research papers, or can be from public resources such as news articles or encyclopedias. We utilize the text corpora associated with the target domain to extract the set of concepts to be inserted into $T$, as explained in Section 4.1.

We extract the sentences containing co-occurrences of the concept pairs linked by relationship types in $S_R$ as part of training data from the source. However, it is unlikely for a whole lot of information to be available about the emergent, target domain concepts in the corpora. Thus as part of the target training data, we restrict BOLT-K to use at least one and at most five labeled instances per relationship type in $T_R$. These would contain co-occurrences of a small number of target concept pairs. We employ data augmentation techniques to enhance this limited amount of training data (Section 4.2). Our BOLT-K approach thus makes use of (i) the abundant labeled relationship information from $S$ and; (ii) the minimal amount of labeled relation information from $T$, to learn relationships between various pairs of concepts for the target domain.

## 4 THE BOLT-K FRAMEWORK

In this section we describe our proposed BOLT-K approach (Figure 1). Among the set of concepts that are to be part of the target ontology $T$, we first identify the pairs of concepts that are likely to be related to each other in $T$, and filter out the remaining spurious pairs. Next, we augment the limited number of available training instances that are labeled with the target relationships from $T_R$. We then use this dataset along with the annotated information from $S$ to learn the relations between the remaining target concepts in $T$. The identified concept pairs for $T$ and the relationship types between them that have been predicted by our model can be used to construct a complete ontology for the target domain $T$.

We now describe each step of our BOLT-K framework in detail.

### 4.1 Filtering Unrelated Target Concept Pairs

This is the first step of our algorithmic pipeline. We acquire the entity and category concepts from the given source ontology structure $S$. We assume for the purpose of simplification, that a list of the concepts that need to be a part of the target ontology $T$ has been provided to us by a domain expert. Alternatively, we can also use emerging entity extraction algorithms [10, 14] to locate and extract concepts from the target text document corpus. This corpus can consist of news reports, laboratory records or research articles authored by domain experts, related to the target topic. It is highly inefficient (quadratic complexity) and unnecessary to consider all

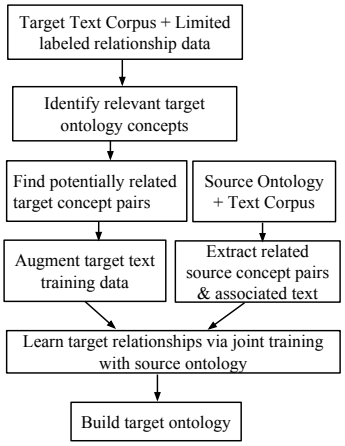


**Figure 1: Pipeline of our BOLT-K framework**

possible pairs of target concepts while learning relationships between them. Hence, our next step is to identify the concept pairs likely to be linked by a relationship and filter out the remaining pairs. Our empirical investigations (Figure 3) show that using only semantic information is insufficient to capture the likelihood of a relationship between pairs of target concepts. Hence, we also incorporate structural information by modeling the target concepts as a graph. We construct a weighted concept graph $T_G$, whose nodes consist of the concepts that are to be a part of the target ontology $T$. We link two nodes by an edge if they have co-occurred together at least once in a document in the target text corpus, within $W$ words of each other. Inspecting a sample of documents in our corpora showed that related concepts often co-occur not in the same sentence, but in consecutive sentences. Since the length of an average sentence is about 10-12 words, we use $W = 25$ to indicate adjacent sentences. The weight of an edge is given by the pointwise mutual information (PMI) between the two concept nodes of the edge:

$$PMI(c_1, c_2) = log\Big(\frac{num(c_1, c_2)}{num(c_1) \cdot num(c_2)}\Big)$$

where $num(c)$ (or $num(c_1, c_2)$) is defined as the number of occurrences of a particular concept (or co-occurrences of a pair of concepts within $W$ words of each other) in the target text corpus.

The PMI metric semantically gives us a good sense of the possibly related target concepts. Nevertheless, it yields a number of false positives which we seek to eliminate, by utilizing additional topological properties of the concept graph. We examine if the local neighborhood of a concept is a sufficient indicator of its potential relationships. However, we empirically find that local structural information is insufficient to account for concept relationships (Figure 3). Therefore, we employ a more global measure for this purpose that takes into account the overall concept graph topology, namely edge-based random walk betweenness centrality [46] (also called current-flow betweenness centrality [3]). For each edge in the weighted graph $T_G$, it measures approximately how often a node is traversed by a random walker going from any node in the network to another. If a concept $c_1$ appears often on random walks from concept $c_2$, then it is likely to be related to $c_2$. Finally, we learn a global threshold based on the betweenness centrality values, and

consider only those target concept pairs as potentially related if their edge betweenness centrality value is above the threshold.

We now have the set of source concept pairs that are related via labels in $S_R$ to each other from the source ontology $S$. We have also obtained the set of target concept pairs that are likely to be related to each other and which will form the target ontology $T$. We now generate a training dataset for both the source and target domains, which will serve as input to the next step of BOLT-K's pipeline, i.e. learning the relationship between the pairs of target concepts. From the source and target document corpora, we extract sentences or blocks of sentences in which potentially related concepts co-occur within $W$ words of each other. Henceforth, for ease of understanding, we call a group of $W$ consecutive words a *sentence*, even though they may physically span more than one sentence of text. We mark the occurrences of the related concepts or entities in each sentence. As mentioned earlier, the target dataset may not have an abundant amount of labeled data available. To account for this, we limit BOLT-K to use at most 5 labeled concept pairs (and hence at most 5 sentences containing them) per target relationship type for training.

## 4.2 Data Augmentation for the Target Domain

A crucial requirement of building a predictive model is the availability of a sufficient amount of labeled training data for the relationship types of the target domain. Hence, to control generalization error and avoid overfitting to the limited training data available, we adopt the technique of *training data augmentation*. It artificially enhances labeled training datasets by transforming the available data items such that the class label properties are preserved. This lends a major advantage to our approach, i.e. the ability to transfer properties from one taxonomic hierarchy to another at no additional annotation cost. Some effective methods of performing text data augmentation with minimal human effort could be to create sentence paraphrases, or to substitute specific words or phrases with likely candidate words (e.g. synonyms). In our work, we augment the relationship-annotated target training sentences by replacing chosen words in them. We want to do this as diversely as possible, such that the syntactic and semantic equivalence between the original and altered sentence is maintained. Inspired by Ratner et al [52], we first identify the noun, verb and adjective terms in the target training sentences using the StanfordCoreNLP part-of-speech tagger [39]. We filter out the terms that do not occur above a learned frequency threshold. Out of these selected terms, we then iteratively sample a term occurring to the left, in between, and to the right of a pair of entities or concepts in each target sentence for substitution. We do not replace more than two terms in a single sentence at a time, to preserve the meaning and grammatical correctness of the modified sentence. We replace the chosen terms in the following three ways.

The first is by constructing an n-gram language model [52]. It is built by recording the frequencies of n-gram occurrences in the source and target corpora, filtering out the less frequent n-grams, and applying Laplace smoothing to the n-gram counts. This model samples words conditioned on the words preceding them. It identifies the n-gram $n_x$ preceding the word or phrase $x$ to be replaced. It then finds from the corpora a list of terms $l_x$, following $n_x$, sorted in descending order based on frequency. It finally replaces $x$ with the term at index $i$ in $l_x$. $i$ is picked based on a geometric distribution $P[i] \sim p^i$, where a more frequent term has a higher probability of being chosen as a substitute [83]. The value of $p$ is fixed at 0.5. The n-gram model falls back to using bigrams (or unigrams), in case the required trigram (or bigram) was filtered out.

For our second technique, we replace the chosen words in each sentence with one of their synonyms from their synset gloss in WordNet [15]. Since synonyms in a gloss are ranked according to how frequently they are observed in natural language, we use a similar geometric distribution as mentioned above to pick a synonym substitute. For our third substitution strategy, we use a pre-trained word2vec [41] word embedding model induced on texts from PubMed, PMC and the English Wikipedia [43]. After obtaining vector representations for each of the terms to be replaced, we substitute them with the terms most similar to their vector representations in the embedding space. We present a comparative evaluation of the three augmentation strategies in Table 3. Once we enhance the labeled dataset for the target domain, we use this data in combination with the labeled examples from the source domain to train the core model of BOLT-K, as described in Section 4.3.

## 4.3 BOLT-K Core Model

We now introduce the core model architecture of BOLT-K. It utilizes human-annotated knowledge on ontological concept relationships from a source domain, and minimally labeled and artificially augmented data from a functionally similar target domain, to learn a hierarchical ontology for the target domain. We build an LSTM-based model with attentive pooling to learn an ontology for the target concepts. We transfer ontological relationship knowledge from the source to the target domain in this model, by sharing the hidden layer representation and some of the model parameters between the two domains. We also combine the objective functions of both domains for effective training. Such models have been used in the literature for transfer learning in various applications [74, 81].

*4.3.1 Base Model Components.* Figure 2 presents an overview of the core model of our proposed BOLT-K framework. It consists of two parts: a *base* model architecture which is shared among both the source and target domains, followed by domain-specific neural network layers. We first describe each of the base model components in detail.

The input to our model is a set of sentences obtained from the source or target text corpora. Each sentence consists of $n$ words $[x_1, x_2, ..., x_n]$. These words include a pair of ontology entities linked by a relation label. Recent literature (e.g. [16, 25, 64, 67, 77]) has seen immense success in transforming words into high-dimensional embeddings for diverse applications. The next layer of our model is thus an embedding layer, which represents every input word $x_i$ as an embedding $d_i$. $d_i$ is formed by concatenating the character-level representation of $x_i$ with its word embedding from a pre-trained word2vec model [43]. This word2vec model has been trained on texts from PubMed, PMC and the English Wikipedia. Words which lack embeddings in the word2vec model are given a random representation. We obtain the character-level representation of each word using a CNN. CNNs have been shown to encode useful morphological information like word prefixes and suffixes from the characters of a word [8, 37, 56]. Our CNN model consists

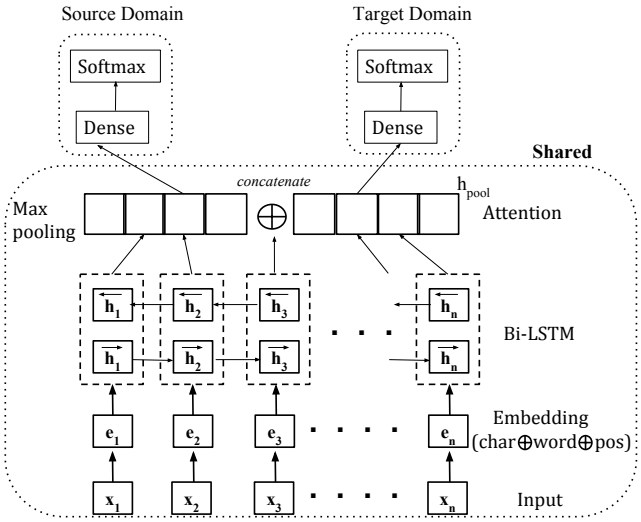


**Figure 2: Proposed BOLT-K model architecture**

of a convolution layer followed by dropout and max pooling. Its input is randomly initialized for each character of a given word $x_i$, and its output is a character-level representation of $x_i$. We form a final embedding $e_i$ by appending two additional position indicators to each $d_i$. These are the normalized word-distances of word $x_i$ from both the related concepts in their respective sentence. The embedding matrix $[e_1, e_2, ..., e_n]$ is updated during model training and serves as input to the next layer, i.e. the bidirectional LSTM.

Bidirectional LSTM [18] based models are used for a variety of sequence modeling tasks where it is often beneficial to utilize both the past and future context. These networks extend the traditional uni-directional LSTM [20] units that only consider past sequential information, by accounting for temporal context information from future time steps. At the core of the LSTM unit is a memory cell controlled by three sigmoidal gates: the input gate $i_t$ deciding whether the unit retains its current input $x_t$ or not, the forget gate $f_t$ to enable the unit to forget its previous memory context $c_{t-1}$, and the output gate $o_t$ controlling the context transferred to the hidden state $h_t$. The recurrences for the LSTM are defined as:

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t * c_{t-1} + i_t * \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t * \tanh(c_t) \end{aligned}$$

where $\sigma$ is the sigmoid function, tanh is the hyperbolic tangent function and $*$ represents the product with the gate value. $W$, $U$ and $b$ are matrices of network parameters to be trained. As shown in Figure 2, the Bi-LSTM layer combines its forward ($\overrightarrow{h_t}$) and backward ($\overleftarrow{h_t}$) sequence contexts using the concatenation operation ($\oplus$). The output $h_t$ of this layer at time step $t$ is given by:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$$

The Bi-LSTM layer generates a sequence of word-level representations $[h_1, h_2, ..., h_n]$ utilizing past and future context, where $h_{j,t}$ denotes the $j$-th element of $h_t$. We perform one-dimensional max pooling to obtain a fixed length vector from the Bi-LSTM output.

$$m_{pool,j} = \max_{1 \le t \le n} [h_{j,t}]$$

The max pooling operation assumes that the important and relevant latent semantic features in the sentence are present at the positions [28] containing the maximum value of $h_t$. However, this might not always be the case. Hence, to focus on words that are crucial in predicting the relationship between entities which might lie anywhere in a sentence, and to give less importance to irrelevant information in the sentences, we utilize an *attention* mechanism. It highlights the important tokens in a given input sequence, which are responsible for performing feature selection for the model as well as for its predictions. Inspired by Yang et al [75], we introduce a word-level attention layer to capture the similarity of a word token with respect to its neighboring context tokens in an input sequence. It assigns weights to the hidden outputs $h_t$ from the Bi-LSTM layer as follows:

$$\begin{aligned} z_{j,t} &= \tanh(W_z h_{j,t} + b_z) \\ \alpha_{j,t} &= \frac{exp(W_a^T z_{j,t})}{\sum_t exp(W_a^T z_{j,t})} \\ att_{pool,j} &= \textstyle\sum_t \alpha_{j,t} h_{j,t} \end{aligned}$$

Here, $W_z$ and $b_z$ are the weight matrix and bias vector respectively associated with the hidden state $h_t$ from the Bi-LSTM layer, and $h_{j,t}$ represents the $j$-th sentence. Their non-linear transformation yields $z_t$, for which $W_a$ is the corresponding weight matrix. $\alpha_{j,t}$ are the normalized attention weights representing token importances from a softmax function at time step $t$. $att_{pool,j}$ is the attention-focused hidden state representation, given by the linear combination of the Bi-LSTM output $h_t$ and the attention weights.

The final output of the base model $h_{pool,j}$ is the concatenation ($\oplus$) of the outputs of the max pooling layer and the attention layer.

$$h_{pool,j} = m_{pool,j} \oplus att_{pool,j}$$

*4.3.2 Knowledge Transfer Components.* All the layers in the base model described in Section 4.3.1, namely the input, embedding, bidirectional LSTM, attention and pooling layers are shared and informed by training data from both the source and target domains. The output from the base model or the shared block in Figure 2 serves as input to two separate fully connected dense layers, followed by two softmax layers, one each for the source and target domains. The softmax function predicts the relationship type between a pair of concepts in an input sentence.

*4.3.3 Training.* We next outline how we transfer information from the source to the target domain by training our model jointly for both domains. We adopt a training procedure similar to that described by Yang et al. [74]. At each iteration, we sample one of the domains from the source and target based on a binomial distribution, where the binomial probability is a tuned parameter. We then optimize the objective function of the chosen domain by training on a sampled batch of labeled instances. The parameters of the shared block are thus updated due to training inputs from both domains, while the fully connected layers are only affected by their corresponding domain. We repeat this procedure until convergence, with early stopping based on the target domain performance.

We used 200-dimensional LSTM units with L2 regularization in our framework. We optimized the cross-entropy error between the true and predicted labels using Adam [27] with gradient clipping. We set the initial learning rate to 0.001 with a decay of 0.05. Dropout [60] was applied to the Bi-LSTM and pooling layers with a probability value of 0.5. Computational support for all our experiments was provided by the Ohio Supercomputer Center [5].

# 5 EVALUATION

## 5.1 Data Collection

We used the Open Biological and Biomedical Ontology (OBO) Foundry [59] and the National Center for Biomedical Ontology portal [44] which are collaborative repositories of science based ontologies, to obtain a family of ontologies related to the bio-medical domain. Each ontology has been curated with the important and relevant concepts associated with its specific sub-domain or subject, as well as the relationship types among the concepts. We also created and used commercial product ontologies from the open-source Web Data Commons project [50]. It contains structured data extracted from the web on different topics. Table 1 shows the statistics of various ontologies we have tested our BOLT-K framework on. These include for each ontology the number of concepts present in it, the number of related concept pairs, the total number of relationship types, and the median number of sentences available for a single relationship type. The first six rows are based on various human diseases and disorders. The next two rows show ontologies of flowering plants (*Angiosperms*) and non-flowering plants (*Gymnosperms*). The final three rows are based on three popular kinds of commercial products that are frequently manufactured, bought and sold, namely, *Earphones, Phones* (mobile phones) and *Television sets*. These ontologies are diverse and contain a total of less than 10K concepts, and less than 20 unique kinds of relationships.

We picked ontologies of sub-domains that are based on related subjects shown in Table 1, and used them interchangeably as both *source* and *target* ontologies to test our framework. For instance, *Dengue* and *Malaria* are both vector-borne diseases transmitted by mosquitoes and share some causes, symptoms and effects, so we use them as a source-target pair. Likewise, *Alzheimers, Multiple Sclerosis, Depression* and *Anxiety* are mental health disorders; *Gymnosperms* and *Angiosperms* are two classes of plant varieties; *Earphones* are an accessory of *Phones*; and *Phones* and *Televisions* are electronic devices sharing some common properties. Hence, we also used these as source-target ontology pairs to test BOLT-K.

As mentioned in Section 3, we require a corpus of text documents containing sentences associated with the concepts in the source and target ontologies $S$ and $T$. To fulfill this requirement for the bio-medical sub-domains, we used a combination of PubMed [4], PubMed Central (PMC) and the English Wikipedia. For an ontology subject $\mathfrak{C}$, we consider all those articles from PubMed, PMC and Wikipedia as part of a text corpus associated with $\mathfrak{C}$ if they contain the term $\mathfrak{C}$ either in their title or abstract. For the last three rows of product hierarchies in Table 1, we constructed a text corpus from the product information and descriptions extracted from the Amazon Product Dataset [19]. This dataset includes information about the numerous commodities sold online on www.amazon.com.

**Table 1: Statistics of various domain ontologies used**

| Ontology sub-domain name | No. of concepts | No. of concept pairs | No. of relations | Median no. of sentences per relation |
|---|---|---|---|---|
| Dengue | 5035 | 5923 | 11 | 6010 |
| Malaria | 2643 | 3556 | 11 | 5070 |
| Alzheimers | 5738 | 5961 | 2 | 9602 |
| Multiple sclerosis | 9036 | 11310 | 2 | 16518 |
| Depressive disorder | 2008 | 4576 | 3 | 3025 |
| Anxiety disorder | 1978 | 4194 | 3 | 3637 |
| Gymnosperms | 539 | 502 | 9 | 716 |
| Angiosperms | 306 | 302 | 10 | 590 |
| Earphones | 115 | 146 | 11 | 28 |
| Phones | 189 | 337 | 16 | 256 |
| Television sets | 72 | 87 | 11 | 8 |

## 5.2 Results

*5.2.1 Identifying Related Concept Pairs.* We first present in Figure 3 the performance of BOLT-K as well as two other baselines on the first step of identifying potentially related concepts for the target ontology, as described in Section 4.1. We consider each ontology in Table 1 as a target ontology.

Our first baseline (yellow bars of *context similarity* in Figure 3) associates a context with each target concept. This is a set of nearby words around the mention of the concept in the target text corpus. If a target concept has multiple mentions (and hence multiple contexts) in the corpus, we pick the context associated with a randomly selected mention. We then compute an embedding for each concept using a normalized term-frequency (*tf*-) weighted sum of the embeddings of its context terms. We use a pre-trained word2vec model [43] to get the term embeddings, ignoring the context terms that do not have embeddings in this model. The *tf*- weights are obtained from the frequencies of occurrence of the chosen context terms, in the available contexts associated with every mention of the target concept. Once we generate an aggregated embedding of each concept, we compute the pairwise cosine similarity between all pairs of concept embeddings. We filter out all those concept pairs that do not have a similarity value above a learned threshold, and consider the remaining pairs as potentially related. Our second baseline (green bars of *2-hop jaccard* in Figure 3) estimates structurally similar concepts as possibly related to each other. It computes the 2-hop Jaccard index between pairs of connected concepts in the concept graph defined in Setion 4.1. For concepts $c_1$ and $c_2$, the Jaccard index is given by the number of neighbor concepts common to both $c_1$ and $c_2$ and reachable from them in 2 hops, divided by the union of the neighbors of $c_1$ and $c_2$. The pink bars of *RWBC* in Figure 3 denote the edge-based random walk betweenness centrality technique used in BOLT-K. The orange portions seen in the last three bars show an overlap between the result numbers of the context-based similarity and the 2-hop Jaccard strategies.

The RWBC measure considers a more global view of the concept graph along with semantic attributes. We find that this causes a marked improvement over merely using either semantic information (the context similarity baseline), or the local structural neighborhood of concepts in the weighted concept network (the Jaccard index baseline). Our random walk betweenness centrality measure

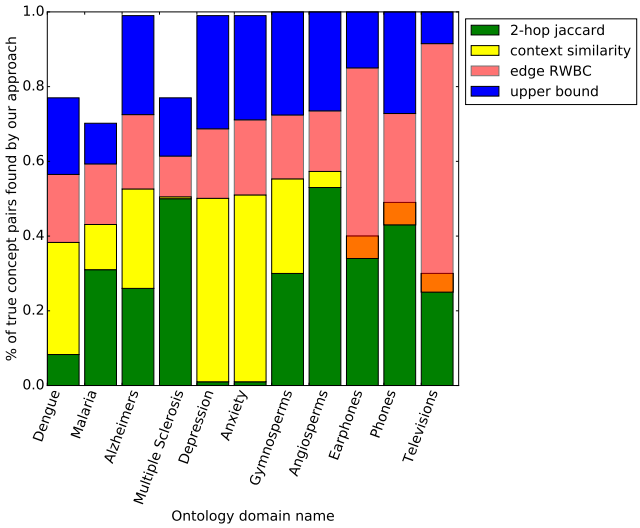


**Figure 3: Performance of multiple approaches on identifying the concepts that are to constitute the target ontology**

correctly identifies as *related*, about 60-75% of the linked concept pairs for the bio-medical ontologies. For the product hierarchies, it identifies more than 70% of the related concept pairs. We note here that every pair of related concepts that is present in the considered ontologies may not occur in the text corpora, i.e. may not be accompanied by textual information. There is thus a limitation on the number of potentially related concept pairs that can be found by our technique. The blue bars show this *upper bound* on the identifiable pairs of linked concepts for each ontology instance. This value is 76-80% for *Dengue*, *Malaria* and *Multiple Sclerosis*, and nearly 100% for the remaining hierarchies. Changing the scale of the result of BOLT-K's edge-based betweenness centrality strategy based on this upper bound, we observe an accuracy of more than 75% in finding potentially related concept pairs for the ontology instances.

*5.2.2 Constructing Domain Ontologies via Knowledge Transfer.* To the best of our knowledge, no existing work learns ontological relationships for a new target subject by levering prior knowledge from a distinct but related source subject. Thus, we compare our method with existing state-of-the-art techniques that only use information from the target domain. We evaluate all approaches on the task of ontological relationship prediction for the target ontology in Table 2. We reiterate that BOLT-K *does not* assume that the source and target ontologies have the same types of relationships. It uses at most 5 relationship-labeled concept pair training examples per target domain. We compare BOLT-K with the following baselines:

(1) PCNN-Att [79]: It employs a piecewise CNN with a sentence-level attention mechanism and distant supervision.
(2) Path-Max [80]: It considers CNNs coupled with probabilistic relation paths learnt from the sentences between entities.
(3) BLSTM-Att [84]: It uses a Bi-LSTM model without pooling, and with an attention mechanism different from ours.
(4) ComplEx [61]: It uses low-rank matrix factorization to learn complex-valued embeddings for entities and relations. We found ComplEx to outperform multiple other knowledge graph completion techniques like TransR [31], TransD [23], DistMult [73] and HolE [47], so we report only its results.
(5) BOLT-K ($t_{sr} = 0.6$) - no Att: This is a variant of our model at a target sub-domain data sampling rate $t_{sr} = 0.6$ (the probability of the model being trained on data from the target sub-domain), on using max pooling without attention.
(6) BOLT-K ($t_{sr} = t_{sr}$): The last five rows of Table 2 are variants of BOLT-K at target data sampling rates $t_{sr} = \{0, 0.4, 0.6, 0.8, 1\}$.

The first column of Table 2 shows the different approaches, and the subsequent columns show the ontology datasets. The $S \rightarrow T$ and $S \leftarrow T$ sub-columns for each column $S \leftrightarrow T$ indicate the direction of knowledge transfer, from ontology $S$ to $T$ and $T$ to $S$ respectively for our approach. The 'arrows' have no significance for the other approaches, merely indicating the ontology for which relations are being learnt ($T$ in case of $S \rightarrow T$ and $S$ in case of $S \leftarrow T$). These results have been computed after applying all augmentation strategies from Section 4.2 to the target training dataset.

We observe that BOLT-K obtains a 5-25% F1-score gain over the baselines in learning relationships for the different ontologies, with the best performance at a target training data sampling rate of 0.6. This reinforces our hypothesis of the utility of levering prior source knowledge in learning ontologies for a newly emergent target domain or sub-domain. BOLT-K is outperformed by ComplEx [61] by about 1% and 4% F1 score points respectively on the *Earphone* $\rightarrow$ *Phone* and the *Phone* $\rightarrow$ *Earphone* product ontology pairs. We believe this is because though these two products are somewhat related, their distinct characteristics may falsely inform our model and detract from its performance. BOLT-K performs better on the product pair of *Phones* and *Televisions* which share relatively more common attributes. It is interesting to observe the difference in performance on interchanging the source and target subject ontologies. For example, there is more value in transferring knowledge from the *Phone* to the *Television* domain, compared to the reverse.

We further note that the overall performance of all approaches on the bio-medical ontologies is significantly better than on the commercial product ontologies. A reason for this could be the quality of textual data available. The text descriptions are often quite generic and repetitive across different product relation types (e.g. *"details about apple iphone 6 16gb - at&t - gold - great condition"*). The variation and uniqueness in context and sentence structure is much lesser than the text data for the bio-medical ontology datasets that come from research, news or encyclopedia articles. Since ComplEx [61] utilizes less information from textual descriptions compared to the other methods, it performs better on the product datasets which have lower quality text input. We also analyze the topological structure of the concept graph (from Section 4.1) for the various ontology datasets. We find multiple well-separated connected components in case of the bio-medical ontologies, signifying a higher separability of their relationship categories. However, the product datasets have few ($\leq 3$) connected components for more than 10 relation categories. This implies that it is harder to distinguish between the product relationship types based on the current information.

*5.2.3 Role of Target Training Data Size.* Augmenting the limited amount of available target training data is a crucial step in our approach. Figure 4 (top) shows the change in F1-score of BOLT-K at the

**Table 2: Baseline F1 scores on the ontology pairs of Dengue ↔ Malaria, Alzheimers ↔ Multiple Sclerosis, Gymnosperms ↔ Angiosperms, Earphones ↔ Phones, and Phones ↔ Televisions. We use at most 5 target training sentences per relation type with data augmentation. $t_{sr}$ is the target sub-domain sampling probability. The first sub-column of every S ↔ T source-target ontology pair denotes knowledge transfer from S → T and the second denotes S ← T (i.e. knowledge transfer from T → S).**

| Approach | Deng ↔ Mal | | Alz ↔ Mult Scl | | Depr ↔ Anxi | | Gymno ↔ Angio | | Earph ↔ Phone | | Phone ↔ TV | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S→T | S←T | S→T | S←T | S→T | S←T | S→T | S←T | S→T | S←T | S→T | S←T |
| PCNN-Att | 0.468 | 0.51 | 0.501 | 0.45 | 0.577 | 0.63 | 0.555 | 0.6 | 0.44 | 0.47 | 0.4 | 0.43 |
| Path-Max | 0.6 | 0.589 | 0.578 | 0.55 | 0.655 | 0.72 | 0.624 | 0.602 | **0.501** | 0.5 | 0.471 | 0.469 |
| BLSTM-Att | 0.601 | 0.62 | 0.505 | 0.49 | 0.567 | 0.6 | 0.592 | 0.63 | 0.43 | 0.45 | 0.41 | 0.399 |
| ComplEx | 0.561 | 0.626 | 0.54 | 0.5 | 0.66 | 0.701 | 0.58 | 0.54 | 0.48 | **0.575** | 0.514 | 0.455 |
| BOLT-K ($t_{sr}$=0.6) - no Att | 0.679 | 0.68 | 0.572 | 0.54 | 0.701 | 0.7 | 0.698 | 0.687 | 0.466 | 0.5 | 0.531 | 0.476 |
| BOLT-K ($t_{sr}$=1) | 0.66 | 0.64 | 0.502 | 0.53 | 0.582 | 0.68 | 0.631 | 0.67 | 0.47 | 0.5 | 0.5 | 0.47 |
| BOLT-K ($t_{sr}$=0.8) | 0.67 | 0.66 | 0.53 | 0.53 | 0.609 | 0.71 | 0.667 | 0.65 | 0.467 | 0.47 | 0.51 | 0.46 |
| BOLT-K ($t_{sr}$=0.4) | 0.644 | 0.63 | 0.545 | 0.5 | 0.65 | 0.66 | 0.63 | 0.62 | 0.44 | 0.45 | 0.46 | 0.44 |
| BOLT-K ($t_{sr}$=0) | 0.527 | 0.51 | 0.499 | 0.48 | 0.576 | 0.6 | 0.55 | 0.58 | 0.398 | 0.401 | 0.41 | 0.399 |
| BOLT-K ($t_{sr}$=0.6) | **0.713** | **0.728** | **0.61** | **0.562** | **0.748** | **0.747** | **0.724** | **0.725** | 0.487 | 0.53 | **0.551** | **0.498** |

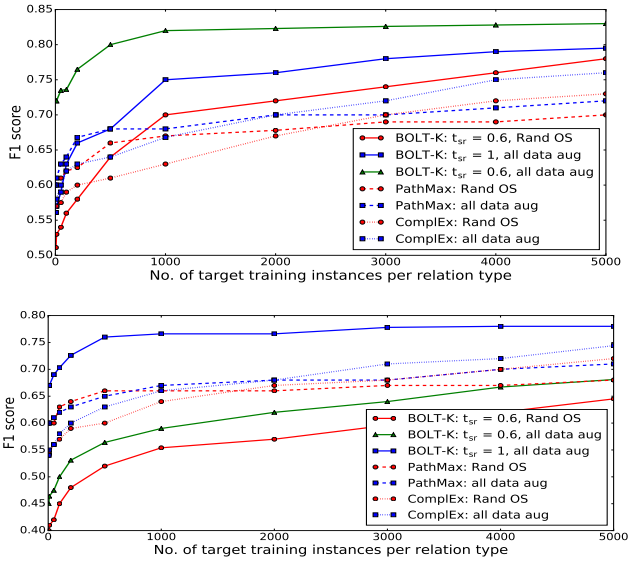


**Figure 4: F1-score vs training data size for Dengue → Malaria (top) and Dengue → Angiosperms (bottom) ontology pairs.**

target data sampling rates $t_{sr} = \{0.6, 1\}$ for the *Dengue → Malaria* ontology pair. We compare BOLT-K to the best performing baselines from Table 2 (*PathMax* [80] and *ComplEx* [61]), as the number of training instances per relation type are increased. We also show in Figure 4 (bottom) an attempt to learn an ontology for *Angiosperms* using information from a dissimilar and unrelated subject, *Dengue*. The solid, thick dashed and thin dashed line plots denote BOLT-K, PathMax and ComplEx respectively. The red plots use only random oversampling to augment the training data, while the blue and green plots use all strategies in Section 4.2. We observe that in case of *Dengue → Malaria* (top plot), as expected, the F1-score of each model rises with the amount of training data it receives. This gain is particularly significant at the left of the x-axis, from 5 to about 1000 training instances per relation category. Employing better data augmentation strategies than random oversampling contributes to the performance. Transferring knowledge from the source ontology lends us a consistent advantage of at least 5%, even at the maximum number of target training instances per relation type. In case of *Dengue → Angiosperms* (bottom plot), BOLT-K performs the best at $t_{sr} = 1$, i.e. without any input from the unrelated source (blue solid line plot). It obtains a 4% gain over ComplEx at the maximum number of training examples per relation class. However, training input from the dissimilar source *Dengue* causes a performance drop (red and green solid line plots).

We next examine in Table 3 the impact of different data augmentation strategies on the performance of BOLT-K, on the *Dengue → Malaria* ontology pair. We observe similar trends on the other datasets as well. The first two rows show the performance using simple random oversampling and Synthetic Minority Oversampling (SMOTE) [6] on the target training examples, followed by the three augmentation strategies described in Section 4.2. The last row of Table 3 shows that these three strategies complement one another. Employing them all together with random oversampling gives us a performance advantage of about 4-8% over any one of them.

*5.2.4 Role of Attention.* Table 2 indicates that attentive pooling lends BOLT-K an F1 score gain of nearly 5%. Further drilling down, Figure 5 shows a heatmap of the attention weight values learned by BOLT-K for sentences belonging to different relation types. The cells of the heatmap contain the sentence word they represent. The background color highlights the importance of a word with respect to its neighboring context, and consequently, how it affects our model's decision. The words in boldface are the concept phrases between which a relation is being predicted. For instance, the first row of the heatmap demonstrates that the words 'refers', 'to', and 'a' (dark background) are crucial in deciding the presence of an *is-a* relation between the concept phrases *virion assembly* and *process of dengue virus*. Similarly, in the fifth row, the words 'begins', 'after' and 'follows' influence the detection of the *preceded-by* relation between the concept phrases *incubation period* and *dengue disease course*. These examples show that BOLT-K has correctly learnt the relevant semantics needed to detect ontological relations between emergent concept phrases.

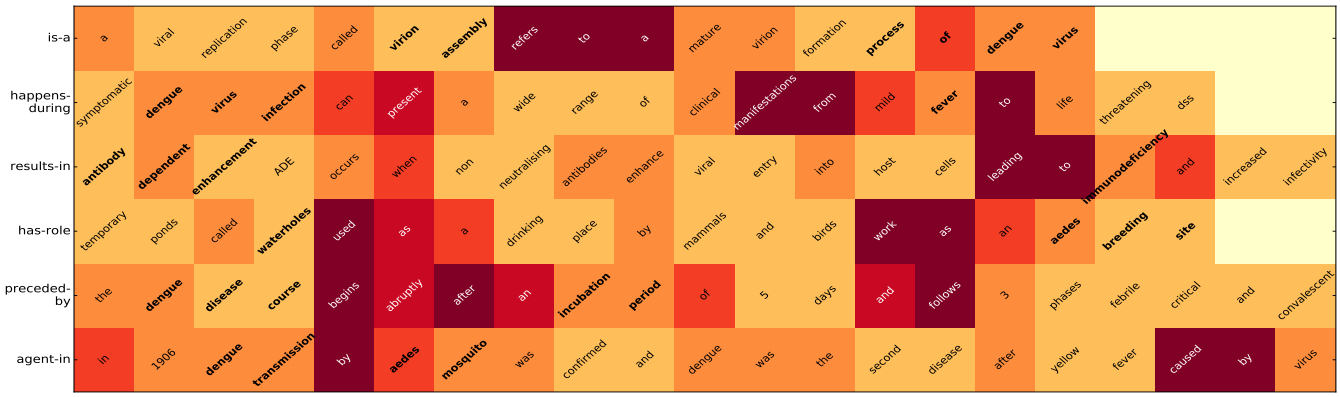


**Figure 5: Visualizing word attention weights for sentences of certain relation types, for the Dengue ontology. The related concept phrase pairs in the sentence are in boldface, and a darker background corresponds to a higher attention weight.**

**Table 3: Assessing various training data augmentation methods for BOLT-K, on the Dengue → Malaria ontology pair**

| Augmentation Technique | Target F1 score |
|---|---|
| Random oversampling (Rand OS) | 0.511 |
| SMOTE oversampling | 0.458 |
| WordNet synonym based replacement + Rand OS | 0.66 |
| Word embedding based replacement + Rand OS | 0.637 |
| Trigram model based replacement + Rand OS | 0.69 |
| WordNet + Word embedding + Trigram + Rand OS | **0.713** |

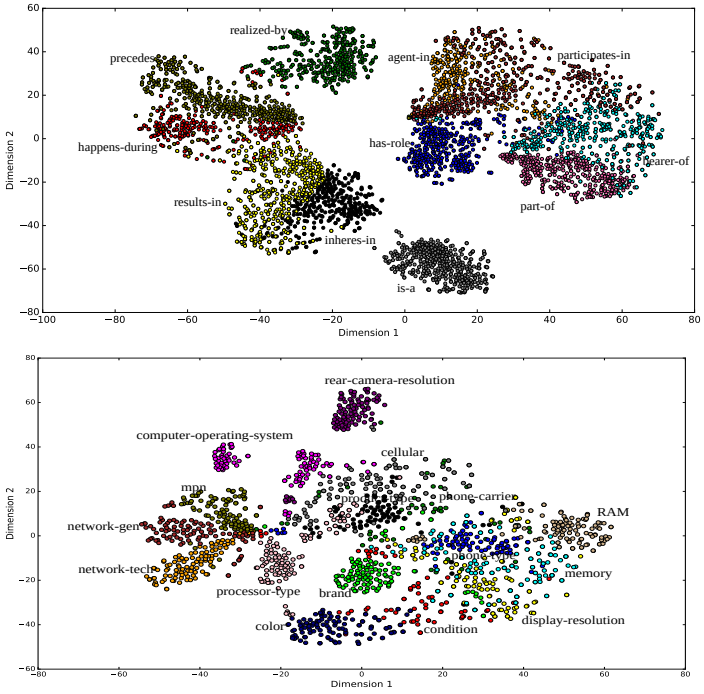


**Figure 6: Visualizing learnt concept pair representations for Dengue → Malaria (top) and Televisions → Phones (bottom).**

## 5.3 Discussion

We now dive deeper into the kind of ontological relationship information that our model can transfer across domains. Figure 6 displays the 2-dimensional t-SNE [38] visualizations of the representations learnt by BOLT-K for the sentence inputs of various relation classes. We observe that the relationships are largely well separable in case of the *Dengue → Malaria* ontology pair (Figure 6 top), with semantically similar relations co-located in their vector space. For example, the relations *participates-in, agent-in, has-role, part-of* and *bearer-of* are situated nearby, and away from the relatively dissimilar relations *results-in* and *inheres-in*. We do not see as clear a segregation between the relation categories for *Televisions → Phones* (Figure 6 bottom). But we do find some relations alike in meaning close by in the vector space too, e.g. *RAM* and *memory*, *network-gen* and *network-tech*.

Table 4 shows the relation categories dominantly mispredicted by our model and the relation type that they are frequently mispredicted as, for different ontology datasets. It also reports for each mispredicted relation category *r*, the percentage of test instances of category *r* that were mispredicted. We find that a common reason for mispredictions is the high similarity in the meaning of certain relationships, due to which they can be used interchangeably in natural language. For instance, BOLT-K is unable to distinguish between the relations *is-a* and *type-of* in case of *Depression ↔ Anxiety*, *participates-in* and *has-role* in case of *Dengue ↔ Malaria* and *preceded-by* and *develops-from* in case of *Gymnosperms ↔ Angiosperms*. We also observe this trend in some cases of the product ontology pairs (last four rows of Table 4), such as *RAM* and *memory*, *brand* and *phone-type*, and *total-size* and *viewable-size*.

As part of further analysis, we seek to understand how well our model can recognize hitherto unseen relation types. For this purpose, we remove or "leave out" one relationship class (i.e. all concept pairs and sentences associated with it) completely from the training dataset of the target sub-domain (and from the training data of the source sub-domain if present). We train our model with

**Table 4: Frequently mispredicted relation types for a set of ontology pairs. For a relation type *r*, the parentheses show the percentage of mispredicted test instances of type *r*.**

| Source → Target | Mispred. rel type: Frequently mispred. as |
|---|---|
| Dengue → Malaria | *part-of* (35%): *is-a*; *inheres-in* (21%): *part-of*; *participates-in* (16%): *has-role*; *agent-in* (7%): *participates-in*; *bearer-of* (6%): *agent-in* |
| Malaria → Dengue | *part-of* (17%): *is-a*; *has-role* (17%): *part-of*; *happens-during* (8%): *part-of*; *precedes* (6%): *happens-during* |
| Alzheimers → Multiple Sclerosis | *type-of* (40%): *has-type*; *has-type* (38%): *type-of* |
| Multiple Sclerosis → Alzheimers | *type-of* (47%): *has-type*; *has-type* (41%): *type-of* |
| Depression → Anxiety | *is-a* (36%): *type-of*; *is-a* (12%): *has-type* |
| Anxiety → Depression | *is-a* (24%): *type-of*; *type-of* (18%): *has-type* |
| Gymnosperms → Angiosperms | *part-of* (19%): *is-a*; *adjacent-to* (20%): *part-of*; *has-part* (21%): *develops-from*; *located-in* (36%): *part-of* |
| Angiosperms → Gymnosperms | *preceded-by* (16%): *develops-from*; *develops-from* (11%): *part-of*; *has-participant* (11%): *part-of*; *part-of* (7%): *is-a* |
| Earphones → Phones | *brand* (42%): *phone-type*; *phone-type* (38%): *brand*; *display-res* (29%): color; *phone-carrier* (25%): *network-gen*; *RAM* (25%): *memory* |
| Phones → Earphones | *model* (44%): *product-type*; *brand* (30%): *model*; *color* (30%): *brand*; *additional-features* (22%): *compatibility*; *tagline* (20%): *brand* |
| Phones → Televisions | *refresh-rate* (39%): *display-type*; *model* (29%): *product-type*; *viewable-size* (29%): *total-size*; *display-res* (23%): *display-type* |
| Televisions → Phones | *RAM* (58%): *memory*; *cellular* (34%): *mpn*; *phone-type* (32%): *brand*; *display-res* (27%): *phone-type*; *cellular* (27%): *phone-carrier* |

**Table 5: Novel relation detection for Dengue → Malaria pair.**

| Relation type *r* left out during training | % of left out concept pairs detected as novel | Novel concept pairs missed. *s*(x%) means x% concept pairs related by *r* were mispredicted as having relation *s*. |
|---|---|---|
| *is-a* | 76% | *part-of(8%), has-role(5%)* |
| *part-of* | 51% | *has-role(19%), is-a(10%)* |
| *happens-during* | 66% | *results-in(11%), part-of(6%)* |
| *precedes* | 70% | *results-in(12%), happens-during(9%)* |
| *results-in* | 61% | *precedes(14%), happens-during(9%)* |
| *has-role* | 39% | *participates-in(21%), agent-in(17%)* |
| *inheres-in* | 58% | *part-of(17%), bearer-of(10%)* |
| *agent-in* | 41% | *participates-in(22%), bearer-of(19%)* |
| *participates-in* | 46% | *has-role(23%), part-of(17%)* |
| *bearer-of* | 64% | *agent-in(13%), part-of(9%)* |
| *realized-by* | 67% | *results-in(9%), precedes(6%)* |

the remaining training data as described in Section 4. We then use the Isolation Forest algorithm [33] with the authors' best case parameter settings to see if BOLT-K can detect concept pairs with the unseen relation type as *novel*, i.e. linked by a novel or unseen relationship. Isolation Forest takes as input the $d$-dimensional features obtained from BOLT-K's penultimate layer before the softmax layer. This novelty detection experiment shows our model's capability in identifying new or unseen relation categories which cannot be easily substituted by another, previously seen relation type.

The results are shown in Table 5, for the *Dengue* → *Malaria* ontology pair. The second column shows the fraction of concept pairs (whose relation class was left out while training) that have been detected as novel. For distinctive relationship types such as *is-a*, we find that more than 75% of the concept pairs which are linked by this relation in the ground truth are predicted as novel. But for left out relations which are not as semantically distinct from the other relationship types, BOLT-K is unable to satisfactorily identify that they were part of an unseen class. For instance, BOLT-K is unable to flag as novel more than 50% of the concept pairs linked by *has-role* and *participates-in*, when they were left out during training. For the concept pairs that were connected by the left-out relationship yet were not detected as novel, we investigated the relationship categories that BOLT-K was predicting them as. The third column of Table 5 shows the two dominant relation types that most concept pairs (linked by the left-out relationship) are mispredicted as belonging to. We found these predictions to be largely logical, when compared with the actual ground truth relationships. For example, 17% of the concept pairs linked by an *inheres-in* relation that was unseen during training were mispredicted as being connected by a *part-of* relation. 21% of the concept pairs connected by a *has-role* relation were mispredicted as having a *participates-in* relation. These mispredictions are semantically plausible due to the similarity in meaning of the misunderstood relationship groups.

## 6 CONCLUSION AND FUTURE WORK

We present BOLT-K, a Bi-LSTM framework with attentive pooling. It identifies concepts relevant to newly emerging subjects or events in various domains from their textual accounts, and automatically learns structured ontologies for them. We bootstrap this process by transferring knowledge from an existing, related sub-domain. We also leverage training data augmentation to accentuate the limited expert-labeled data available for these emergent sub-domains, at no further annotation cost. We extensively evaluate BOLT-K on real-life bio-medical and commercial product ontologies.

BOLT-K currently involves manual intervention to define functionally or logically related domains for knowledge transfer during ontology learning. In future, we plan to automate this task. Since the meaning and validity of concepts and their relationships may change over time, we seek to enrich BOLT-K by learning concept and relation representations that can effectively capture their evolving dynamics. This can enable the prediction of time instances at which concepts or relations are likely to appear, disappear or reappear with respect to an ontology snapshot. We also aim to improve BOLT-K's scalability so it can handle larger and deeper ontologies.

## ACKNOWLEDGMENTS

This work is supported by the National Science Foundation grant EAR-1520870. The authors would also like to thank Xin Luna Dong for suggestions on product datasets. All content presented represents the opinion of the authors, and is not necessarily endorsed by their sponsors.

## REFERENCES


[1] Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6, Nov (2005), 1817–1853.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*. 2787–2795.

[3] Ulrik Brandes and Daniel Fleischer. 2005. Centrality measures based on current flow. In *Annual Symposium on Theoretical Aspects of Computer Science*. Springer, 533–544.

[4] Kathi Canese and Sarah Weis. 2013. PubMed: the bibliographic database. (2013).

[5] Ohio Supercomputer Center. 1987. Ohio Supercomputer Center. http://osc.edu/ark:/19495/f5s1ph73. (1987).

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[7] Yu Chen and Mohammed J Zaki. 2017. Kate: K-competitive autoencoder for text. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 85–94.

[8] Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional LSTM-CNNs. *arXiv preprint arXiv:1511.08308* (2015).

[9] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, Aug (2011), 2493–2537.

[10] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 140–147.

[11] Xin Luna Dong. 2018. Challenges and innovations in building a product knowledge graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2869–2869.

[12] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. 2014. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*. 766–774.

[13] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440* (2017).

[14] Michael Färber, Achim Rettinger, and Boulos El Asmar. 2016. On emerging entity detection. In *European Knowledge Acquisition Workshop*. Springer, 223–238.

[15] WordNet Fellbaum. 1998. An Electronic Lexical Database (Language, Speech, and Communication). (1998).

[16] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *ACL Vol 1 Long Papers*.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[18] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6645–6649.

[19] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 507–517.

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[21] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955* (2017).

[22] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059* (2018).

[23] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 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)*, Vol. 1. 687–696.

[24] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*. 198–202.

[25] A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE CVPR*.

[26] J-D Kim, Tomoko Ohta, Yuka Tateisi, and JunâĂŹichi Tsujii. 2003. GENIA corpus: a semantically annotated corpus for bio-textmining. *Bioinformatics* 19, suppl_1 (2003), i180–i182.

[27] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[28] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification.. In *AAAI Conference on Artificial Intelligence*, Vol. 333. 2267–2273.

[29] Ig-hoon Lee, Suekyung Lee, Taehee Lee, Sang-goo Lee, Dongkyu Kim, Jonghoon Chun, Hyunja Lee, and Junho Shim. 2005. Practical issues for building a product ontology system. In *Proceedings of the International Workshop on Data Engineering Issues in E-Commerce*. IEEE, 16–25.

[30] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115* (2017).

[31] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion.. In *AAAI Conference on Artificial Intelligence*, Vol. 15. 2181–2187.

[32] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 2124–2133.

[33] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 413–422.

[34] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*. 136–144.

[35] Xinghua Lu, Bin Zheng, Atulya Velivelli, and ChengXiang Zhai. 2006. Enhancing text categorization with semantic-enriched representation and training data augmentation. *Journal of the American Medical Informatics Association* 13, 5 (2006), 526–535.

[36] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 745–754.

[37] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016).

[38] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[39] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60.

[40] Suyu Mei, Wang Fei, and Shuigeng Zhou. 2011. Gene ontology based transfer learning for protein subcellular localization. *BMC bioinformatics* 12, 1 (2011), 44.

[41] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[42] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 1003–1011.

[43] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*. 39–43.

[44] Mark A Musen, Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Christopher G Chute, Margaret-Anne Story, Barry Smith, and NCBO team. 2011. The national center for biomedical ontology. *Journal of the American Medical Informatics Association* 19, 2 (2011), 190–195.

[45] Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI: International Joint Conference on Artificial Intelligence*.

[46] Mark EJ Newman. 2005. A measure of betweenness centrality based on random walks. *Social networks* (2005).

[47] Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, and others. 2016. Holographic Embeddings of Knowledge Graphs.. In *AAAI Conference on Artificial Intelligence*, Vol. 2. 3–2.

[48] Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. *arXiv preprint arXiv:1603.00786* (2016).

[49] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *arXiv preprint arXiv:1708.03743* (2017).

[50] Petar Petrovski, Anna Primpeli, Robert Meusel, and Christian Bizer. 2016. The WDC gold standards for product feature extraction and product matching. In *International Conference on Electronic Commerce and Web Technologies*. Springer, 73–86.

[51] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1003–1012.

[52] Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. 2017. Learning to Compose Domain-Specific Transformations for Data Augmentation. In *Advances in Neural Information Processing Systems*. 3239–3249.

[53] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 148–163.

[54] Itsumi Saito, Jun Suzuki, Kyosuke Nishida, Kugatsu Sadamitsu, Satoshi Kobashikawa, Ryo Masumura, Yuji Matsumoto, and Junji Tomita. 2017. Improving Neural Text Normalization with Data Augmentation at Character-and Morphological Levels. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 257–262.

[55] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*. 1163–1171.

[56] Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1818–1826.

[57] Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580* (2015).

[58] Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. *arXiv preprint arXiv:1705.00316* (2017).

[59] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, and others. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* 25, 11 (2007), 1251.

[60] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[61] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*. 2071–2080.

[62] Luu Tuan, Yi Tay, Siu Hui, and See Ng. 2016. Learning Term Embeddings for Taxonomic Relation Identification Using Dynamic Weighting Neural Network. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[63] Nikhita Vedula, Patrick K Nicholson, Deepak Ajwani, Sourav Dutta, Alessandra Sala, and Srinivasan Parthasarathy. 2018. Enriching Taxonomies With Functional Domain Knowledge. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 745–754.

[64] Nikhita Vedula and Srinivasan Parthasarathy. 2017. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*. ACM, 127–136.

[65] Nikhita Vedula, Srinivasan Parthasarathy, and Valerie L Shalin. 2016. Predicting trust relations among users in a social network. On the roles of influence, cohesion and valence. *Proceedings of ACM SKGKDD WISDOM* (2016).

[66] Nikhita Vedula, Srinivasan Parthasarathy, and Valerie L Shalin. 2017. Predicting trust relations within a social network: A case study on emergency response. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 53–62.

[67] Nikhita Vedula, Wei Sun, Hyunhwan Lee, Harsh Gupta, Mitsunori Ogihara, Joseph Johnson, Gang Ren, and Srinivasan Parthasarathy. 2017. Multimodal Content Analysis for Effective Advertisements on YouTube. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1123–1128.

[68] Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics* 39, 3 (2013), 665–707.

[69] VG Vydiswaran, ChengXiang Zhai, and Dan Roth. 2011. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 974–982.

[70] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes.. In *AAAI Conference on Artificial Intelligence*, Vol. 14. 1112–1119.

[71] John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847* (2017).

[72] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation Learning of Knowledge Graphs with Entity Descriptions.. In *AAAI Conference on Artificial Intelligence*. 2659–2665.

[73] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* (2014).

[74] Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345* (2017).

[75] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.

[76] Jianxing Yu, Zheng-Jun Zha, Meng Wang, Kai Wang, and Tat-Seng Chua. 2011. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 140–150.

[77] Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning Term Embeddings for Hypernymy Identification.. In *IJCAI*.

[78] Jianbo Yuan, Han Guo, Zhiwei Jin, Hongxia Jin, Xianchao Zhang, and Jiebo Luo. 2017. One-shot learning for fine-grained relation extraction via convolutional siamese neural network. In *2017 IEEE International Conference on Big Data*. IEEE, 2194–2199.

[79] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1753–1762.

[80] Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2016. Incorporating relation paths in neural relation extraction. *arXiv preprint arXiv:1609.07479* (2016).

[81] Jing Zhang, Wanqing Li, and Philip Ogunbona. 2017. Transfer learning for cross-dataset recognition: a survey. *arXiv preprint arXiv:1705.04396* (2017).

[82] Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710* (2015).

[83] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*. 649–657.

[84] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 207–212.