Improving Zero-Shot Cross-lingual Transfer for Multilingual Question Answering over Knowledge Graph

Yucheng Zhou[1*], **Xiubo Geng**[2], **Tao Shen**[3], **Wenqiang Zhang**[1†], **Daxin Jiang**[2†]
[1]Fudan University, Shanghai, China
[2]Microsoft, Beijing, China
[3]Australian AI Institute, School of CS, FEIT, University of Technology Sydney
{yczhou18,wqzhang}@fudan.edu.cn, tao.shen@student.uts.edu.au
{xigeng,djiang}@microsoft.com

Abstract

Multilingual question answering over knowledge graph (KGQA) aims to derive answers from a knowledge graph (KG) for questions in multiple languages. To be widely applicable, we focus on its zero-shot transfer setting. That is, we can only access training data in a high-resource language, while need to answer multilingual questions without any labeled data in target languages. A straightforward approach is resorting to pre-trained multilingual models (e.g., mBERT) for cross-lingual transfer, but there is a still significant gap of KGQA performance between source and target languages. In this paper, we exploit unsupervised bilingual lexicon induction (BLI) to map training questions in source language into those in target language as augmented training data, which circumvents language inconsistency between training and inference. Furthermore, we propose an adversarial learning strategy to alleviate syntax-disorder of the augmented data, making the model incline to both language- and syntax-independence. Consequently, our model narrows the gap in zero-shot cross-lingual transfer. Experiments on two multilingual KGQA datasets with 11 zero-resource languages verify its effectiveness.

1 Introduction

With the advance of large-scale human-curated knowledge graphs (KG), e.g., DBpedia (Auer et al., 2007) and Freebase (Bollacker et al., 2008), question answering over knowledge graph (KGQA) has become a crucial natural language processing (NLP) task to answer factoid questions. It has been integrated into real-world applications like search engines and personal assistants, so it attracts more attention from both academia and industry (Liang et al., 2017; Hu et al., 2018; Shen et al., 2019).

Recently, a rising demand of KGQA systems is to answer the multilingual questions, motivating us to focus on multilingual KGQA. However, building a large-scale KG, as well as annotating QA data, is costly for each new language, not to mention many minority languages with a few native annotators. Therefore, we adopt a zero-shot cross-lingual transfer setting – a KGQA model is developed to perform inference on multilingual questions with the only access to training data and associated KG in a high-resource language (e.g., English).

Providing the success of pre-trained monolingual encoders (Peters et al., 2018; Liu et al., 2019), some works (e.g., mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020)) pre-train a Transformer encoder (Vaswani et al., 2017) on large-scale non-parallel multilingual corpora in a self-supervised manner. Then given an NLP task, a general paradigm for zero-shot cross-lingual transfer is to fine-tune a pre-trained multilingual encoder on the data in a data-rich (*source*) language. And the fine-tuned model is generalizable enough to perform inference in other low-resource (*target*) languages with surprising quality of prediction. This paradigm can be adapted to KGQA to build symbolic logical forms (e.g., query graph (Yih et al., 2015)) for KG query. However, it is witnessed that there is a considerable KGQA performance gap between source and target languages, which is consistent with the empirical results on a wide range of other tasks by prior works (Conneau et al., 2020).

To bridge the gap, translation approaches are proven effective on multilingual benchmarks (Hu et al., 2020; Liang et al., 2020). As a way of data augmentation, they perform source-to-target translation to obtain multilingual training data. Further with advanced techniques (Cui et al., 2019; Fang et al., 2020), they achieve state-of-the-art effectiveness. But these approaches rely heavily on a well-performing translator. The translator is not always available especially for a minority language since its training requires a large volume of parallel bilingual corpus. Therefore, to be applicable to

[*]Work is done during internship at Microsoft.
[†]Corresponding authors.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5822–5834
June 6–11, 2021.

more languages, we assume that neither translators nor parallel corpora are available in this work.

In this paper, to adapt the translation approaches in our zero-resource scenario, we naturally propose to replace the full-supervised machine translator with unsupervised bilingual lexicon induction (BLI) for word-level translation. Specifically, as in prior works (Lample et al., 2018b; Artetxe et al., 2018), a BLI model is first trained on non-parallel bilingual corpora. Then, via bilingual word alignments in BLI, we map the training questions in source language into those in target languages to obtain augmented multilingual training data. Consequently, even simply learning a KGQA model on the augmented data can circumvent language inconsistency between training and inference and thus bridge the performance gap in zero-shot cross-lingual transfer. To explain why BLI is competent, it is observed that KGQA mainly involves phrase-level semantics (Berant et al., 2013). Compared to other tasks depending on sentence-level contextualization, KGQA is insensitive to long-term dependency but benefits from the language consistency.

Moreover, we propose an adversarial strategy to mitigate the syntax-disorder caused by BLI. Specifically, we present a discriminator on top of the encoder, which is trained to distinguish whether the input is a grammatical question in source language or a BLI-translated one in target language. Meanwhile, jointly with KGQA goal, the encoder is fine-tuned to fool the discriminator so that the questions' representations are both language- and syntax-agnostic. So the trained KGQA model is robust to syntax-disorder and becomes insensitive to the question language, leading to superior performance on multilingual KGQA.

Experiments conducted on two multilingual KGQA datasets with 11 zero-resource languages verify the effectiveness of our approach.

2 KGQA Task Definition

We give a background of monolingual KGQA, followed by multilingual KGQA and its data format.

Monolingual KGQA. A knowledge graph $\mathcal{G}$ is comprised of a set of directed triples (h, p, t), where $h \in \mathcal{E}$ denotes a head entity, $t \in \mathcal{E} \bigcup \mathcal{L}$ denotes a tail entity or literal value, and $p \in \mathcal{P}$ denote a predicate between h and t. KGQA aims at generating answers for a natural language question q based on $\mathcal{G}$. Usually a model $\mathcal{M}$ first parses the question q into an intermediate logical form, which is then transformed into a SPARQL query, and the answer is derived by executing the SPARQL query on $\mathcal{G}$. An example is shown in Figure 1: the question in the bottom, intermediate logical form in the upper right and the corresponding SPARQL query in the top. Following Maheshwari et al. (2019), we take a restricted subset of λ-calculus – query graph, as the intermediate logical form. Typically, a query graph consists of four types of nodes: grounded entity(s) (in rounded rectangle), existential variable(s) "*?y*" (in circle), a lambda variable "*?x*" (in shaded circle), and an aggregation function (in diamond).

Considering entity-linking is a standalone system and there are many tools, we assume grounded entities in a question are given. This avoids uncertainty caused by entity-linking, and facilitates us to focus on the query graph construction process.

Multilingual KGQA. We focus on a zero-shot cross-lingual transfer setting of KGQA. That is, we only have a labeled dataset $\mathcal{D}^{src} = \{(q_l^{src}, s_l^{src})_{l=1}^{N}\}$, as well as the associated knowledge graph $\mathcal{G}$, in a high-resource language src, where q_l^{src} and s_l^{src} denote a natural language question and a formal query, respectively. We will omit subscript l of example index in $\mathcal{D}^{src}$. Multilingual KGQA is to learn a model $\mathcal{M}$ which can answer questions q^{tgt} in multiple target languages tgt. A recent baseline is to fine-tune pre-trained multilingual models (e.g. mBERT) in src and directly perform inference in tgt.

3 Methodology

This section starts with a base framework for monolingual KGQA, followed by our proposed multilingual solutions. Lastly, details about training and inference are elaborated.

3.1 Base Monolingual Framework

Following Maheshwari et al. (2019), we present a base pipeline framework as in Figure 1 to construct query graphs. It consists of three modules: 1) inferential chain ranking, 2) type constraint ranking, and 3) aggregator classification.

Inferential Chain Ranking. An inferential chain (IC) refers to a sequence of directed predicate from a grounded entity to lambda variable $?x$. Given an entity e grounded from the question q, we first search its chain candidates $\mathcal{C}^e = (c_1^e, \ldots, c_n^e)$ by exploring legitimate predicate sequences starting from e in $\mathcal{G}$. Following previous works (Yih

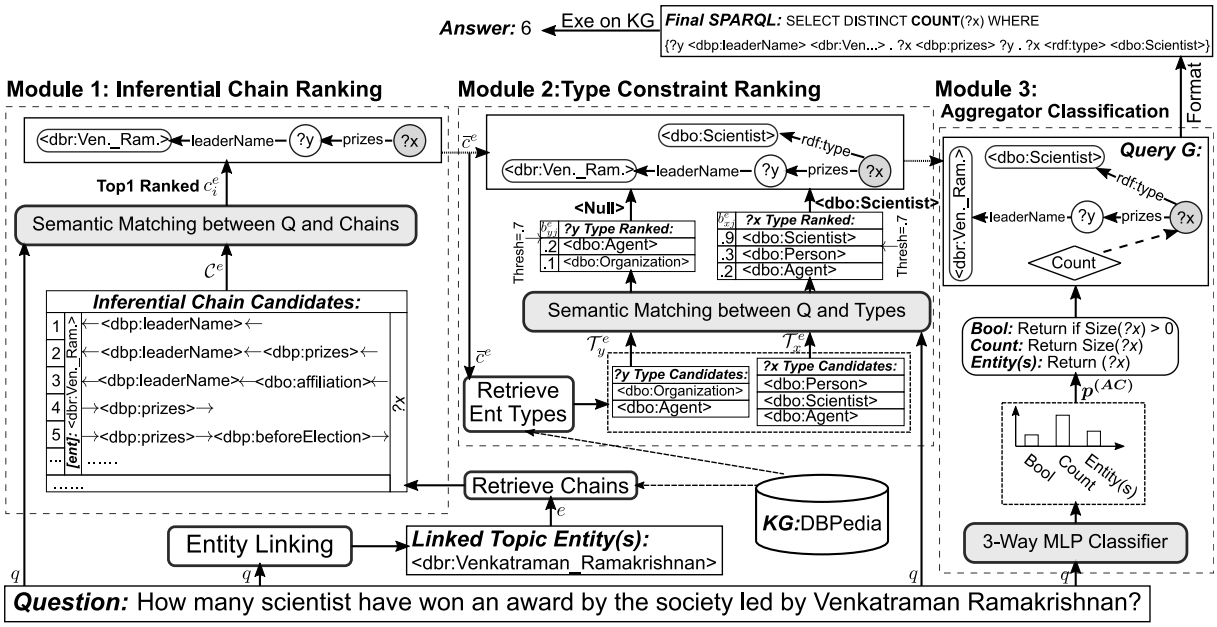

Figure 1: Base framework for monolingual KG, consisting of three modules to construct a query graph.

et al., 2015; Maheshwari et al., 2019), we fetch the chains whose length ≤ 2. For example, as in the middle left of Figure 1, chain candidates are generated from the entity "*<dbr:Ven.-Ram.>*" within 2-hop on $\mathcal{G}$. Then, a model is presented to measure the semantic relatedness between the question q and each candidate of inferential chain c_i^e, i.e.,

$$a_i^e = \text{SemMatch}(q, c_i^e; \theta^{(IC)}), \forall i = 1, \ldots, n, \quad (1)$$

where a_i^e is a score for their relatedness, and $\theta^{(IC)}$-parameterized $\text{SemMatch}(\cdot)$ can be any model for pairwise relatedness, such as Co-Attention network (Chen et al., 2019) and BERT-based Matching (Devlin et al., 2019). Finally, the resulting of this module is the top-1 ranked inferential chain, i.e.,

$$\bar{c}^e = \arg\max_{c_i^e}(a_i^e, \ \forall i = 1, \ldots, n). \quad (2)$$

Note, if there are multiple grounded entities in q, we will predict an inferential chain for each entity.

Type Constraint Ranking. Type constraints (TC) refer to the entity types specified in the question for each variable on an inferential chain. They can be used to disambiguate the entities and thus boost KGQA performance. For example, answer entity(s) to the example question in Figure 1 are constrained by type *Scientist*. Hence, type constraint ranking is proposed to capture such information, which is also achieved by a semantic matching model. Specifically, given the resulting inferential chain $\bar{c}^e$, we first enumerate type candidates $\mathcal{T}_y^e = \{t_{y1}^e, \ldots\}$ for the existential variable and $\mathcal{T}_x^e = \{t_{x1}^e, \ldots\}$ for the lambda variable. Then, because there is scarcely overlap of gold type constraints between the two variables, a single semantic matching model is adequate for both. Thus, we define the model to derive relatedness scores as

$$b_{*j}^e = \text{SemMatch}(q, t_{*j}^e; \theta^{(TC)}), \quad (3)$$
$$\text{where, } \forall * \in \{y, x\}, \text{and } \forall j = 1, \ldots$$

Finally, we get the type constraints for existential and lambda variable with a threshold $\gamma^{(thresh)}$, i.e.,

$$\bar{\mathcal{T}}_*^e = \{t_{*j}^e | b_{*j}^e > \gamma^{(thresh)}, \ \forall j = 1, \ldots\}. \quad (4)$$

Aggregator Classification Given several answer formats in the dataset, aggregator classification (AC) is presented to distinguish the format among *Bool*, *Count* and *Entity(s)*. The principle of each is detailed in the middle right of Figure 1. Formally, a simple text classifier can satisfy, i.e.,

$$\boldsymbol{p}^{(AC)} = \text{Classifier}(q; \theta^{(AC)}) \in \mathbb{R}^3, \quad (5)$$

where the $\text{Classifier}(\cdot)$ is composed of a contextualized encoder, a pooler and an MLP with softmax.

Once the above is completed, their results can compose a query graph, which is transformed into SPARQL and then executed on $\mathcal{G}$ for the answer.

3.2 Proposed Multilingual KGQA Approach

Built upon the base framework detailed before, we extend it with a multilingual inference capability,

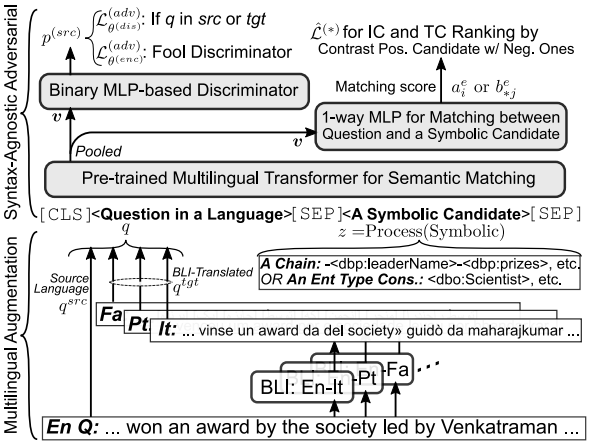

Figure 2: Syntax-agnostic semantic matching between a BLI-augmented multilingual q and its symbolic candidates.

i.e., multilingual KGQA. We are in line with a recent popular zero-shot transfer paradigm (Conneau et al., 2020; Fang et al., 2020) that: a pre-trained multilingual encoder is only fine-tuned in src, and a translation-based data augmentation technique is integrated to narrow the performance gap between src and tgt. To emphasize the gap in KGQA, 65% F1 score in English (src) vs. 54% in Italian (tgt) is observed by mBERT zero-shot transfer in our pipeline without any multilingual augmenting.

Distinct from prior works in this paradigm requiring well-trained translators, we propose a fully unsupervised way for wide applicability with neither tgt KGQA data nor src-tgt parallel corpora. It is natural to resort to bilingual lexicon induction (BLI) with unsupervised training and acceptable word-level translating quality. In the following, we first present a BLI-based augmentation for multilingual training data, followed by our adaptation of the monolingual base framework (§ 3.1) to the augmented data. Finally, we propose an adversarial learning strategy coupled with BLI-based augmentation for robust cross-lingual transfer. An illustration of our proposed semantic matching model with symbolic candidates is in Figure 2.

3.2.1 BLI-based Multilingual Augmentation

We leverage the BLI model by Lample et al. (2018b). First, it pre-trains monolingual word embeddings $\boldsymbol{U}^{src} \in \mathbb{R}^{d\times|\mathcal{V}^{src}|}$ and $\boldsymbol{U}^{tgt} \in \mathbb{R}^{d\times|\mathcal{V}^{tgt}|}$ in src and tgt respectively. Then, it learns a linear transformation to unsupervisedly align the word embeddings in two languages to one space, i.e.,

$$\bar{\boldsymbol{W}} = \underset{\boldsymbol{W}\in M_d(\mathbb{R})}{\arg\min} \sum \text{Distance}(\boldsymbol{W}\boldsymbol{U}^{src}_{:,k}, \boldsymbol{U}^{tgt}_{:,l}). \quad (6)$$

The unsupervised alignment between k-th src word and l-th tgt word is captured by adversarial learning, and $\text{Distance}(\cdot)$ is implemented by cross-domain similarity local scaling (CSLS). Please refer to (Lample et al., 2018b) for its details.

Based on the BLI model, we can build a word-by-word translator, $\text{BLI}^{(trans)}_{src\to tgt}$, from src to arbitrary tgt, as long as its monolingual corpus is available. Note, when performing word-level translation, we also employ CSLS to mitigate the hubness problem and find the most likely alignment. Then, we translate each question q^{src} in $\mathcal{D}^{src}$ to other languages:

$$q^{tgt} = \text{BLI}^{(trans)}_{src\to tgt}(q^{src}), \quad (7)$$

where src denotes English (en) in our experiments while tgt can be one of 11 other languages, such Farsi (fa), Italian (it), etc. Consequently, q^{tgt} is the augmented multilingual data for model training.

Remark: Although BLI provides multilingual data, open questions still remain. 1) *Why is BLI competent here:* It is observed KGQA mainly involves word-/phrase-level semantics of symbolic candidates, rather than sentence-level one in most other NLP tasks. As the Module 1 and 2 in Figure 1, the matching only involves morphological similarity (e.g., *scientist* vs. <*dbo:Scientist*>), synonym (e.g., *won an award* vs. <*dbp:prizes*>), etc. Thus, KGQA is less sensitive to long-term context than other tasks. This has been leveraged by Berant et al. (2013) to propose a phrase matching model for monolingual KGQA. 2) *Will BLI lead to error propagation:* Since BLI model achieves a high Precision@10 but a relatively low Precision@1, wrong translation and the corresponding ground truth are semantically similar. Intuitively, their word embeddings are spatially close to each other, so wrong word-level translation is equivalent to applying tiny noise to word embeddings, which hardly leads to error propagation when robust pre-trained Transformer-based encoder is used.

3.2.2 Multilingual Models

Symbolic Candidate Processing. For an inferential chain, we enrich each predicate on the chain by 1) transforming each camel-represented phrase into sequence-formatted words 2) prefixing +/- for directional information, and 3) concatenating top-frequent types in local closed-world assumptions (Krompaß et al., 2015). For a type constraint, we simply transform each camel-represented phrase into sequence-formatted words. In the following,

we denote the text of a processed symbolic candidate as z no matter it is a chain or type.

Multilingual Semantic Matching Model. As detailed in §3.1, both inferential chain ranking and type constraint ranking modules are built upon a semantic matching model between the question q and a symbolic candidate z. Note, z is always in src while q can be in either src or BLI-translated tgt. Following the common practice, we first concatenate q and z with special tokens (Devlin et al., 2019), which is passed into a pre-trained multilingual Transformer encoder, i.e.,

$$\boldsymbol{v} = \text{Pool}(\text{Transformer}(text)), \quad (8)$$
$$\text{where, } text = (\texttt{[CLS]}, q, \texttt{[SEP]}, z, \texttt{[SEP]}).$$

$\text{Pool}(\cdot)$ denotes using the contextualized embedding of `[CLS]` to represent the entire input. In this paper, the encoder is alternative between mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). Lastly, a 1-way multi-layer perceptron (MLP) built upon $\boldsymbol{v}$ is presented to calculate the matching score in Eq.(1) or Eq.(3).

Multilingual Classification Model. As detailed in §3.1, a text classification model is required to identify aggregator. To fit into our zero-resource multilingual scenario, the model, consisting of a pre-trained multilingual encoder and an MLP-based predicting layer, can be directly fine-tuned on the augmented questions, i.e., q^{src} and q^{tgt}.

3.2.3 Syntax-agnostic Adversarial Strategy

Although training the KGQA model on BLI-augmented multilingual data circumvents language inconsistency, it inevitably introduces syntax disorder and grammatical problem, which could hurt the performance. We thus present an adversarial strategy in pair with BLI-augmented data to push the Transformer encoder deriving language- and syntax-independent representations. Formally, a discriminator is built upon the single vector representation $\boldsymbol{v}$ produced by the Transformer encoder:

$$p^{(src)} = \text{Sigmoid}(\text{MLP}(\boldsymbol{v}; \theta^{(dis)})), \quad (9)$$

where $p^{(src)}$ is the probability of the question in source. The discriminator is trained to minimize

$$\mathcal{L}^{(adv)}_{\theta^{(dis)}} = -\mathbb{I}_{(src)} \log p^{(src)} - \mathbb{I}_{(tgt)} \log(1 - p^{(src)}). \quad (10)$$

On the contrary, the Transformer encoder is learned to fool by minimizing an adversarial loss, i.e.,

$$\mathcal{L}^{(adv)}_{\theta^{(enc)}} = -\mathbb{I}_{(tgt)} \log p^{(src)}. \quad (11)$$

$\mathbb{I}_{(tgt)}$ denotes if the question in BLI-translated tgt, and $\theta^{(enc)}$ is encoder's parameters in each module.

3.3 Training

Before constructing the objectives, we conduct uniform negative sampling for the two ranking models with the maximum negative number limited to 100.

First, gold labels of a q for the three modules stem from the formal query s^{src}. A margin-based hinge loss is defined for inferential chain ranking:

$$\hat{\mathcal{L}}^{(IC)} = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \frac{1}{|\mathcal{N}|} \sum\nolimits_{i=1}^{|\mathcal{N}|} (\lambda - \tilde{a}^e + \hat{a}^e_i), \quad (12)$$

where, $\mathcal{D}$ is the augmented dataset, $\mathcal{N}$ is a set of negative chains, $\tilde{a}^e$ is derived from the gold chain and $\hat{a}^e_i$ is derived from a negative chain. Similarly, the loss defined for type constraint ranking is

$$\hat{\mathcal{L}}^{(TC)} = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \frac{1}{2|\mathcal{N}|} \sum_{*\in\{y,x\}} \sum_{j=1}^{|\mathcal{N}|} (\lambda - \tilde{b}^e_* + \hat{b}^e_{*j}).$$

Lastly, the loss of aggregator classification is

$$\hat{\mathcal{L}}^{(AC)} = -\frac{1}{|\mathcal{D}|} \sum\nolimits_{\mathcal{D}} \log \boldsymbol{p}^{(AC)}_{[i=\tilde{g}]}, \quad (13)$$

where $\boldsymbol{p}^{(AC)}_{[i=\tilde{g}]}$ denotes probability corresponding to gold aggregator class.

During training, the adversarial loss is added to the loss function of each module to compose the final training objective, i.e.,

$$\mathcal{L}^{(*)} = \hat{\mathcal{L}}^{(*)} + \alpha \mathcal{L}^{(adv)}_{\theta^{(enc)}}, * \in \{IC, TC, AC\}. \quad (14)$$

3.4 Inference Algorithm

As in Algorithm 1, we provide a detailed procedure for model inference in target language.

We also provide an explanation of query graph in Figure 1. As the example query graph shown in the right of the figure: a topic entity is first grounded as $e =$"*<dbr:Ven.-Ram>*" in rounded rectangle, an existential variable in circle denotes intermediate entity set $?y = \{h|(h, \textit{leaderName}, e)\}$, a lambda variable in shaded circle denotes the answer entity set $?x = \{h|(h, \textit{prizes}, e) \wedge \forall e \in ?y\}$, and an aggregator COUNT is finally applied to $?x$ that is constrained by entity type "*<dbo:Scientist>*". Note that, the existential variable can not exist if only 1-hop relation is expressed in a question, and if multiple topic entities are grounded, multiple "*?x*" will be merged by intersection.

Algorithm 1 Inference in Target Language.

Require: : A q in tgt and its grounded topic entities $\mathcal{E}^q$; KG $\mathcal{G}$; Models $\theta^{(IC)}$, $\theta^{(TC)}$, $\theta^{(AC)}$

Search the chain candidates $\mathcal{C}^e$ on $\mathcal{G}$, $\forall e \in \mathcal{E}^q$
Rank each $\mathcal{C}^e$ by Eq.(1), and keep top-3 in $\mathcal{C}^e$
$\mathcal{C}^e \leftarrow \{c^e | c^e \in \mathcal{C}^e \wedge \text{Size}(?x \in c^e) > 0\}$
$\bar{c}^e \leftarrow$ *Null*
if $\text{Size}(\mathcal{C}^e) > 0$ **then**
  $\bar{c}^e \leftarrow$ the top1 inferential chain in $\mathcal{C}^e$
end if
Merge chains $\{\bar{c}^e | \forall e \in \mathcal{E}^q \wedge \bar{c}^e \; is \; not \; Null\}$
Rank type constraint candidates by Eq.(3) and apply the top-1 constraint w/ score $> \gamma^{(thresh)}$
Generate SPARQL and execute on $\mathcal{G}$ for answer entity set $\mathcal{A}$
Identify the aggregator for q by Eq.(5)
$\mathcal{A} \leftarrow \text{Aggregate}(\mathcal{A})$ by following Figure 1
return $\mathcal{A}$;

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the proposed approach on two datasets, LC-QuAD (Trivedi et al., 2017) and QALD-multilingual (Usbeck et al., 2018), both of which contain questions with corresponding SPARQL queries over DBpedia[1]. DBpedia is a large-scale knowledge graph extracted from Wikipedia pages with 6 million/60 thousands/13 billion entities/predicates/triples in the English edition.

LC-QuAD. LC-QuAD is a large-scale complex question answering dataset, which contains 5000 English question-SPARQL pairs[2]. We follow the official split with 1000 questions in the test set, and further split the original training set into training/valid with 3500/500 questions. To evaluate the effectiveness of multilingual KGQA, questions in the test set are translated into 10 languages (fa, de, ro, it, ru, fr, nl, es, hi, pt)[3] using Google Translator[4].

QALD-multilingual. QALD is a series of evaluation campaigns on question answering over linked data[5]. We collect all multilingual questions along with their SPARQL queries from QALD4 to QALD9 and filter out some out-of-scope ones[6]. There are overall 429 distinct question-SPARQL pairs and most are expressed in 12 languages (en, fa, de, ro, it, ru, fr, nl, es, hi_IN, pt, pt_BR). Considering the small size of this dataset, we take all QALD-multilingual questions as test set, and use the training data of LC-QuAD for model training.

Evaluation Metrics. We adopt two widely-used metrics as following (Maheshwari et al., 2019), i.e., inferential chain accuracy (ICA) and macro F1 score. The former is used to measure the accuracy (i.e., Precision@1) of inferential chain model, and defined as the percent of correctly-predicted inferential chains. The macro F1 score is used to measure the performance of final answers. Please refer to (Maheshwari et al., 2019) for the details.

4.2 Experimental Setting

We evaluate our approach with 2 multilingual encoding models, i.e. $\text{mBERT}_{\text{base}}$ and $\text{XLM-R}_{\text{base}}$. The embedding and hidden size in both models are set to 768. We use Adam optimizer (Kingma and Ba, 2015) to optimize the KGQA loss with the learning rate of 5×10^{-5} and a linear warm-up (Vaswani et al., 2017). The maximum training epoch, warm-up epoch, and batch size are set to 35, 3, and 32. The discriminator is trained along with each module's objective, with α set to 5×10^{-4} for learning to fool. The discriminator is optimized via the Adam optimizer with a learning rate of 5×10^{-5}. $\gamma^{(thresh)}$ for the type constraint model is set to 0.7. We follow (Maheshwari et al., 2019) and use the same values for other parameters in model training.

4.3 Main Results

We compare our approach with a natural, widely-used baseline, which fine-tunes a pre-trained multilingual model (e.g., mBERT, XLM-R) on source language, and then directly apply it to target languages. The comparison on QALD-multilingual and LC-QuAD with mBERT are reported in Table 1 and 2 respectively. It is showed that our approach outperforms the baseline significantly on both datasets for all languages. ICA is improved by 1%-4%, and 2.9% on average on the QALD dataset. The improvement on LC-QuAD is even larger, i.e., averaged ICA and F1 score of all languages are increased by around 7% and 4% respectively. Notably, with the BLI-augmented data and syntax-

[1] We use the 2016-10 version, which can be downloaded at https://wiki.dbpedia.org/downloads-2016-10.
[2] https://github.com/AskNowQA/LC-QuAD.
[3] https://github.com/yczhou001/Multilingual-KBQA-Dataset/tree/main/LC-QuAD.
[4] https://translate.google.com/.
[5] https://github.com/ag-sc/QALD.
[6] https://github.com/yczhou001/Multilingual-KBQA-Dataset/tree/main/QALD.

ICA	en	fa	de	ro	it	ru	fr	nl	es	hi_IN	pt	pt_BR	Avg	Avg w/o en
Baseline	80.7	76.0	77.8	76.8	76.5	80.4	76.9	78.5	77.6	79.3	80.9	86.3	79.0	78.8
Ours	83.7	77.6	80.5	79.2	80.5	83.1	80.3	80.5	81.7	82.5	85.3	87.4	81.9	81.7
Lift	+3.1	+1.6	+2.8	+2.5	+4.0	+2.7	+3.4	+2.0	+4.0	+3.2	+4.4	+1.1	+2.9	+2.9
F1	**en**	**fa**	**de**	**ro**	**it**	**ru**	**fr**	**nl**	**es**	**hi_IN**	**pt**	**pt_BR**	**Avg**	**Avg w/o en**
Baseline	65.0	58.0	60.8	60.2	53.7	60.5	59.8	64.3	55.2	59.3	60.5	70.0	60.6	60.2
Ours	66.7	60.0	62.2	62.1	57.7	63.5	63.6	65.9	58.8	62.6	63.5	70.0	63.0	62.7
Lift	+1.7	+2.0	+1.4	+2.0	+4.0	+3.0	+3.8	+1.7	+3.7	+3.2	+3.1	+0.0	+2.5	+2.5

Table 1: Comparison on QALD-multilingual using mBERT.

ICA	en	fa	de	ro	it	ru	fr	nl	es	hi_IN	pt	Avg	Avg w/o en
Baseline	87.0	83.8	88.3	86.1	86.0	86.0	86.9	87.2	88.2	83.7	86.6	86.3	86.3
Ours	94.7	91.7	93.3	93.1	93.2	92.7	93.1	94.2	94.1	92.6	93.4	93.3	93.2
Lift	+7.7	+7.9	+5.0	+7.0	+7.3	+6.7	+6.2	+7.0	+5.9	+9.0	+6.8	+6.9	+6.9
F1	**en**	**fa**	**de**	**ro**	**it**	**ru**	**fr**	**nl**	**es**	**hi_IN**	**pt**	**Avg**	**Avg w/o en**
Baseline	80.1	66.6	78.3	68.9	69.1	71.1	69.5	75.8	72.9	66.5	69.3	71.6	70.8
Ours	85.5	71.7	82.4	72.6	72.3	74.5	73.2	80.9	76.1	71.9	74.0	75.9	74.9
Lift	+5.4	+5.1	+4.1	+3.6	+3.2	+3.4	+3.6	+5.1	+3.2	+5.5	+4.7	+4.3	+4.2

Table 2: Comparison on LC-QuAD-multilingual using mBERT.

agnostic adversarial learning, the performance of source-language (i.e., English) questions are also increased by a large margin, i.e., F1 score increases from 65% to 66.7% on QALD, and from 80% to 85% on LC-QuAD. We also evaluate the propose approach using XLM-R as the multilingual encoder. The comparison on QALD-multilingual is shown in Table 3. We can observe similar improvements as in mBERT, where both averaged ICA and F1 score are increased by around 1%, verifying the effectiveness of our proposed approach.

4.4 Ablation Study

Our approach consists of two important components, BLI-based data augmentation and a syntax-agnostic learning strategy. We conduct an ablation study to investigate the effect of each component. Table 4 reports the averaged results of all target-languages on QALD-multilingual and LC-QuAD-multilingual. From the table we can see that, with BLI-based data augmentation, our approach increases the ICA score on QALD by 1.7%, and the syntax-agnostic adversarial learning further improves it by 1.2%. Similar improvements are observed on LC-QuAD, which verifies the effectiveness of both components in our approach.

4.5 Analysis

Impact of BLI Accuracy. We assess the impact of BLI accuracy on five Romance languages (i.e. it, fr, es, pt, and ro) by injecting noise into BLI results. Specifically, when mapping source-language words into a target language via BLI, we randomly replace translated words with wrong ones with a probability of p (10%, 20%, 30%, 40%, and 50%). The averaged performance of our approach on the five languages is reported in Figure 3. It is observed, with more noise added, the performance of our approach drops, which is in accordance with intuition. But even when 50% of the translated words are noisy, our method still outperforms the baseline model. For example, it is superior than the baseline by 1% in terms of ICA with 50% noise, showing the robustness of our approach.

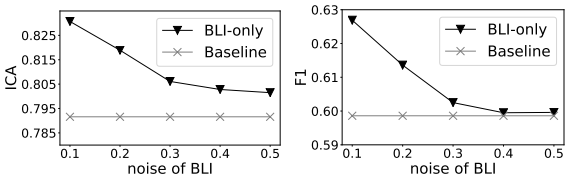

Figure 3: Impact of BLI Accuracy in our approach. The x-axis represents the percentage of noise we inject into BLI results, while y-axis represents the performance in terms of ICA in Figure (left) and F1 score in Figure (right).

Deep Dive into Adversarial Learning. We take the inferential chain ranking model as an example, and take a deep dive into the impact of syntax-agnostic adversarial learning. The adversarial learning involves a discriminator to distinguish whether a question is grammatical or syntax-disorder, and an inferential chain ranking model to identify the gold chain. Their loss values, i.e., $\mathcal{L}^{(dis)}_{\theta^{(dis)}}$ and

ICA	en	fa	de	ro	it	ru	fr	nl	es	hi_IN	pt	pt_BR	Avg	Avg w/o en
Baseline (XLM-R base)	81.5	76.9	75.6	77.7	76.7	80.9	76.5	78.8	77.4	80.2	80.4	84.2	78.9	78.7
Ours (XLM-R base)	84.0	78.1	77.5	79.0	77.1	80.9	77.8	79.4	78.1	81.3	80.9	85.3	79.9	79.6
Lift	+2.5	+1.2	+1.9	+1.4	+0.4	+0.0	+1.3	+0.7	+0.7	+1.2	+0.5	+1.1	+1.1	+0.9
F1	**en**	**fa**	**de**	**ro**	**it**	**ru**	**fr**	**nl**	**es**	**hi_IN**	**pt**	**pt_BR**	**Avg**	**Avg w/o en**
Baseline (XLM-R base)	63.4	57.1	54.7	58.8	50.1	59.4	56.3	61.3	51.2	59.2	57.5	66.1	57.9	57.4
Ours (XLM-R base)	64.6	57.6	56.1	61.4	50.9	59.4	58.2	62.1	52.2	60.6	57.4	66.1	58.9	58.4
Lift	+1.2	+0.5	+1.4	+2.6	+0.8	+0.0	+1.9	+0.8	+1.0	+1.3	-0.1	+0.0	+1.0	+0.9

Table 3: Comparison on QALD-multilingual using XLM-R.

	QALD		LC-QuAD	
Avg w/o en	ICA	F1	ICA	F1
BLI-only	80.5	60.9	91.7	74.2
BLI-only vs. Baseline	+1.7	+0.7	+5.5	+3.4
BLI+Adv.	81.7	62.7	93.2	74.9
BLI+Adv. vs. BLI-only	+1.2	+1.8	+1.4	+0.7

Table 4: Ablation study. "BLI-only vs. Baseline" represents the effect of BLI. "BLI+Adv. vs. BLI-only" represents effect of syntax-agnostic adversarial learning.

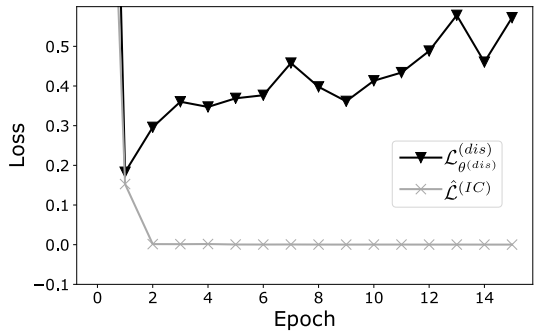

Figure 4: Adversarial losses on validation set w.r.t different epochs in training phase.

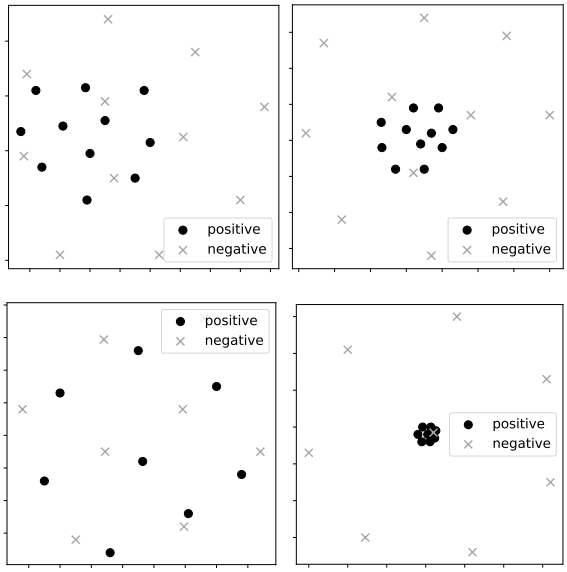

Figure 5: Case study via t-SNE visualization. Different points in a graph represents different languages for the same question. (Upper left) and (upper right) show embeddings of baseline and our approach for the question "what is the population of Cairo?". (Lower left) and (lower right) show embeddings of baseline and our approach for the question "which species does an elephant belong?".

$\hat{\mathcal{L}}^{(IC)}$, are plot in Figure 4. We can see that the classification loss of the discriminator quickly drops and then slowly goes up, indicating that the discriminator gets good performance and then it is fooled later by the language-/syntax-agnostic embeddings generated by mBERT. Meanwhile, the inferential ranking loss drops quickly and stays very small in following epochs, showing that when mBERT is generating syntax-agnostic embeddings, it also supports the inferential chain ranking very well.

4.6 Case Study

We take several examples of inferential chain ranking to show how our approach works. We use t-SNE (Maaten and Hinton, 2008) to map the embedding of a question-chain pair into a two-dimensional data point. A question in a specific language is paired with its golden inferential chain and top-1 ranked negative candidate. Figure 5 compares the baseline with our approach for two questions. Positive and negative examples of the same question in different languages are plot in the same figure. We can see that the baseline model can not distinguish positive inferential chains from negative ones well, while our approach can learn a language-agnostic representation that focuses more on ranking inferential chain candidates.

5 Related Work

There are mainly two categories of approaches to handle monolingual question answering over knowledge graph (KGQA) task. (1) *Information retrieval-based* approaches align a question with its answer candidates in the same semantic space,

where the candidates usually stem from KG neighbors of the topic entity detected in the questions (Bordes et al., 2014b,a; Dong et al., 2015; Jain, 2016; Xu et al., 2016; Hao et al., 2017; Chen et al., 2019). (2) *Semantic parsing-based* approaches first translate a question into the corresponding logical form, e.g., program (Guo et al., 2018; Shen et al., 2019) or query graph (Yih et al., 2015; Jia and Liang, 2016; Xiao et al., 2016; Dong and Lapata, 2016; Liang et al., 2017; Dong and Lapata, 2018; Maheshwari et al., 2019), and then execute the logical form over KG to derive the final answer. Note a logical form is usually composed of a series of grammars or operators pre-defined by experts. This paper is in line with the second category to generate query graph for KG execution. To the best of our knowledge, there are only few works targeting multilingual KGQA (Hakimov et al., 2017; Veyseh, 2016), which rely on extensive multilingual training data with hand-crafted features while are inapplicable to the zero-shot transfer scenario. So we adopt the pipeline by Maheshwari et al. (2019) for monolingual scenario as our base model but update the encoders with the Transformer (Vaswani et al., 2017) to strengthen their expressive power and facilitate recent pre-trained multilingual initializations.

Given task-specific data in a source language, cross-lingual models are trained to perform inference in target languages in a low- or zero-resource scenario. Typically, cross-lingual models are proposed in two paradigms. 1) *Universal encoding*-based paradigm represents multilingual natural language text into language-agnostic embeddings the same semantic space. Early works focus on aligning multilingual word embedding (Mikolov et al., 2013; Faruqui and Dyer, 2014; Xu et al., 2018), while recent efforts are mainly made on large-scale pre-trained multilingual encoder, such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), Unicoder (Huang et al., 2019a), XLM-R (Conneau et al., 2020), InfoXLM (Chi et al., 2020), and ALM (Yang et al., 2020). They can perform zero-shot cross-lingual transfer by training in the source language while directly inference in target language. 2) *translation*-based paradigm employs well-trained machine translators to map the training or test examples in source language to those in target translation. Recent common practice tends to leverage the second paradigm to generate multilingual data to narrows the zero-shot cross-lingual performance gap in the first paradigm, which leads to state-of-the-art results on several cross-lingual benchmarks. In contrast, we consider a zero-resource scenario where translators are unavailable and we thus resort to unsupervised BLI in light of KGQA's characteristics.

As a branch of universal encoding at word level, bilingual lexicon induction (BLI) (a.k.a cross-lingual word embedding – CLWE) is learned to align bilingual word embeddings in the same space, where the embeddings are pre-trained on monolingual corpora and the alignment is trained in either a (semi-)supervised or unsupervised manner (Smith et al., 2017; Lample et al., 2018b; Artetxe et al., 2018, 2019; Huang et al., 2019b; Patra et al., 2019; Karan et al., 2020; Zhao et al., 2020; Ren et al., 2020). To alleviate "hubness" problem (Dinu and Baroni, 2015) in BLI, alternatives of the distance measurement are proposed to substitute nearest neighbor (NN) during the alignment, such as inverted-softmax (Smith et al., 2017) and CSLS (Lample et al., 2018b). In addition to building bilingual dictionary via word-level translation, a well-trained BLI model can serve as a weak baseline of sentence-level translation (Lample et al., 2018a), a seed model for unsupervised translation (Lample et al., 2018a) or a bilingual variant of copy mechanism in summarization (Zhu et al., 2020).

Moreover, adversarial training is usually integrated into cross-lingual models for language-agnostic representation learning, such as unsupervised BLI (Lample et al., 2018b; Zhang et al., 2017), unsupervised translation (Lample et al., 2018a), cross-Lingual sequence labeling (Kim et al., 2017; Huang et al., 2019c) and cross-Lingual classification (Dong et al., 2020). In contrast, our adversarial strategy not only considers language-agnostic representations but also aims at making the model insensitive to syntax-disorder and thus competent in zero-resource scenario.

6 Conclusion

We propose a novel approach for zero-shot cross-lingual transfer in multilingual KGQA, which augments training data by bilingual lexicon induction, and leverages a syntax-agnostic adversarial learning strategy to alleviate the syntax-disorder problem caused by BLI. Experimental results on two multilingual KGQA datasets in 11 zero-resource languages verify its effectiveness.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 789–798. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5002–5007. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014a. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 615–620. ACL.

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014b. Open question answering with weakly supervised embedding models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, volume 8724 of *Lecture Notes in Computer Science*, pages 165–180. Springer.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2019. Bidirectional attentive memory networks for question answering over knowledge bases. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2913–2923. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *CoRR*, abs/2007.07834.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1586–1595. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Georgiana Dinu and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 731–742. Association for Computational Linguistics.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 260–269. The Association for Computer Linguistics.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard de Melo. 2020. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1541–1544. ACM.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2020. FILTER: an enhanced fusion method for cross-lingual language understanding. *CoRR*, abs/2009.05166.

Manaal Faruqui and Chris Dyer. 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, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 462–471. The Association for Computer Linguistics.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-action: Conversational question answering over a large-scale knowledge base. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2946–2955.

Sherzod Hakimov, Soufian Jebbara, and Philipp Cimiano. 2017. Amuse: multilingual semantic parsing for question answering over linked data. In *International Semantic Web Conference*, pages 329–346. Springer.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 221–231. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Sen Hu, Lei Zou, and Xinbo Zhang. 2018. A state-transition framework to answer complex questions over knowledge base. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2098–2108.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019a. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2485–2494. Association for Computational Linguistics.

Jiaji Huang, Qiang Qiu, and Kenneth Church. 2019b. Hubless nearest neighbor search for bilingual lexicon induction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4072–4080. Association for Computational Linguistics.

Lifu Huang, Heng Ji, and Jonathan May. 2019c. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3823–3833. Association for Computational Linguistics.

Sarthak Jain. 2016. Question answering over knowledge base using factual memory networks. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 109–115. The Association for Computational Linguistics.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Mladen Karan, Ivan Vulic, Anna Korhonen, and Goran Glavas. 2020. Classification-based self-learning for weakly supervised bilingual lexicon induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6915–6922. Association for Computational Linguistics.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2832–2838. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Denis Krompaß, Stephan Baier, and Volker Tresp. 2015. Type-constrained representation learning in knowledge graphs. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, volume 9366 of *Lecture Notes in Computer Science*, pages 640–655. Springer.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Chen Liang, Jonathan Berant, Quoc V. Le, Kenneth D. Forbus, and Ni Lao. 2017. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 23–33. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *CoRR*, abs/2004.01401.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Gaurav Maheshwari, Priyansh Trivedi, Denis Lukovnikov, Nilesh Chakraborty, Asja Fischer, and Jens Lehmann. 2019. Learning to rank query graphs for complex question answering over knowledge graphs. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*, volume 11778 of *Lecture Notes in Computer Science*, pages 487–504. Springer.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 184–193. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Shuo Ren, Shujie Liu, Ming Zhou, and Shuai Ma. 2020. A graph-based coarse-to-fine method for unsupervised bilingual lexicon induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3476–3485. Association for Computational Linguistics.

Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-task learning for conversational question answering over a large-scale knowledge base. *arXiv preprint arXiv:1910.05069*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. Lc-quad: A corpus for complex question answering over knowledge graphs. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II*, volume 10588 of *Lecture Notes in Computer Science*, pages 210–218. Springer.

Ricardo Usbeck, Ria Hari Gusmita, Axel-Cyrille Ngonga Ngomo, and Muhammad Saleem. 2018. 9th challenge on question answering over linked data (QALD-9) (invited paper). In *Joint proceedings of the 4th Workshop on Semantic Deep*

Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, California, United States of America, October 8th - 9th, 2018, volume 2241 of *CEUR Workshop Proceedings*, pages 58–64. CEUR-WS.org.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Amir Pouran Ben Veyseh. 2016. Cross-lingual question answering using common semantic space. In *Proceedings of TextGraphs-10: the workshop on graph-based methods for natural language processing*, pages 15–19.

Chunyang Xiao, Marc Dymetman, and Claire Gardent. 2016. Sequence-based structured prediction for semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2465–2474. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9386–9393. AAAI Press.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1321–1331. The Association for Computer Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1959–1970. Association for Computational Linguistics.

Xu Zhao, Zihao Wang, Yong Zhang, and Hao Wu. 2020. A relaxed matching procedure for unsupervised BLI. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3036–3041. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1309–1321. Association for Computational Linguistics.