Cross-language Learning with Adversarial Neural Networks: Application to Community Question Answering

Shafiq Joty, Preslav Nakov, Lluís Màrquez and **Israa Jaradat**
ALT Research Group
Qatar Computing Research Institute, HBKU
{sjoty, pnakov, lmarquez, ijaradat}@hbku.edu.qa

Abstract

We address the problem of cross-language adaptation for question-question similarity reranking in community question answering, with the objective to port a system trained on one input language to another input language given labeled training data for the first language and only unlabeled data for the second language. In particular, we propose to use adversarial training of neural networks to learn high-level features that are discriminative for the main learning task, and at the same time are invariant across the input languages. The evaluation results show sizable improvements for our cross-language adversarial neural network (CLANN) model over a strong non-adversarial system.

1 Introduction

Developing natural language processing (NLP) systems that can work indistinctly with different input languages is a challenging task; yet, such a setup is useful for many real-world applications. One expensive solution is to annotate data for each input language and then to train a separate system for each one. Another option, which can be also costly, is to translate the input, e.g., using machine translation (MT), and then to work monolingually in the target language (Hartrumpf et al., 2008; Lin and Kuo, 2010; Ture and Boschee, 2016). However, the machine-translated text can be of low quality, might lose some input signal, e.g., it can alter sentiment (Mohammad et al., 2016), or may not be really needed (Bouma et al., 2008; Pouran Ben Veyseh, 2016). Using a unified cross-language representation of the input is a third, less costly option, which allows any combination of input languages during both training and testing.

In this paper, we take this last approach, i.e., combining languages during both training and testing, and we study the problem of question-question similarity reranking in community Question Answering (cQA), when the input question can be either in English or in Arabic, and the questions it is compared to are always in English. We start with a simple language-independent representation based on cross-language word embeddings, which we input into a feed-forward multilayer neural network to classify pairs of questions, (English, English) or (Arabic, English), regarding their similarity.

Furthermore, we explore the question of whether *adversarial* training can be used to improve the performance of the network when we have some *unlabeled* examples in the target language. In particular, we adapt the Domain Adversarial Neural Network model from (Ganin et al., 2016), which was originally used for domain adaptation, to our cross-language setting. To the best of our knowledge, this is novel for cross-language question-question similarity reranking, as well as for natural language processing (NLP) in general; moreover, we are not aware of any previous work on *cross-language* question reranking for community Question Answering.

In our setup, the basic task-solving network is paired with another network that shares the internal representation of the input and tries to decide whether the input example comes from the source (English) or from the target (Arabic) language. The training of this language discriminator network is *adversarial* with respect to the shared layers by using gradient reversal during backpropagation, which makes the training to *maximize* the loss of the discriminator rather than to minimize it. The main idea is to learn a high-level abstract representation that is discriminative for the main classification task, but is invariant across the input languages.

We apply this method to an extension of the SemEval-2016 Task 3, subtask B benchmark dataset for question-question similarity reranking (Nakov et al., 2016b). In particular, we hired professional translators to translate the original English questions to Arabic, and we further collected additional unlabeled questions in English, which we also got translated into Arabic. We show that using the unlabeled data for adversarial training allows us to improve the results by a sizable margin in both directions, i.e., when training on English and adapting the system with the Arabic unlabeled data, and vice versa. Moreover, the resulting performance is comparable to the best monolingual English systems at SemEval. We also compare our unsupervised model to a semi-supervised model, where we have some labeled data for the target language.

The remainder of this paper is organized as follows: Section 2 discusses some related work. Section 3 introduces our model for adversarial training for cross-language problems. Section 4 describes the experimental setup. Section 5 presents the evaluation results. Finally, Section 6 concludes and points to possible directions for future work.

2 Related Work

Below we discuss three relevant research lines: (*a*) adversarial training, (*b*) question-question similarity, and (*c*) cross-language learning.

Adversarial training of neural networks has shown a big impact recently, especially in areas such as computer vision, where generative unsupervised models have proved capable of synthesizing new images (Goodfellow et al., 2014; Radford et al., 2016; Makhzani et al., 2016). One crucial challenge in adversarial training is to find the right balance between the two components: the generator and the adversarial discriminator. Thus, several methods have been proposed recently to stabilize training (Metz et al., 2017; Arjovsky et al., 2017). Adversarial training has also been successful in training predictive models. More relevant to our work is the work of Ganin et al. (2016), who proposed domain adversarial neural networks (DANN) to learn discriminative but at the same time domain-invariant representations, with domain adaptation as a target. Here, we use adversarial training to learn task-specific representations in a *cross-language* setting, which is novel for this task, to the best of our knowledge.

Question-question similarity was part of Task 3 on cQA at SemEval-2016/2017 (Nakov et al., 2016b, 2017); there was also a similar subtask as part of SemEval-2016 Task 1 on Semantic Textual Similarity (Agirre et al., 2016). Question-question similarity is an important problem with application to question recommendation, question duplicate detection, community question answering, and question answering in general. Typically, it has been addressed using a variety of textual similarity measures. Some work has paid attention to modeling the question topic, which can be done explicitly, e.g., using a graph of topic terms (Cao et al., 2008), or implicitly, e.g., using LDA-based topic language model that matches the questions not only at the term level but also at the topic level (Zhang et al., 2014). Another important aspect is syntactic structure, e.g., Wang et al. (2009) proposed a retrieval model for finding similar questions based on the similarity of syntactic trees, and Da San Martino et al. (2016) used syntactic kernels. Yet another emerging approach is to use neural networks, e.g., dos Santos et al. (2015) used convolutional neural networks (CNNs), Romeo et al. (2016) used long short-term memory (LSTMs) networks with neural attention to select the important part when comparing two questions, and Lei et al. (2016) used a combined recurrent–convolutional model to map questions to continuous semantic representations. Finally, translation models have been popular for question-question similarity (Jeon et al., 2005; Zhou et al., 2011). Unlike that work, here we are interested in *cross-language adaptation* for question-question similarity reranking. The problem was studied in (Martino et al., 2017) using cross-language kernels and deep neural networks; however, they used no adversarial training.

Cross-language Question Answering was the topic of several challenges, e.g., at CLEF 2008 (Forner et al., 2008), at NTCIR-8 (Mitamura et al., 2010), and at BOLT (Soboroff et al., 2016). Cross-language QA methods typically use machine translation directly or adapt MT models to the QA setting (Echihabi and Marcu, 2003; Soricut and Brill, 2006; Riezler et al., 2007; Hartrumpf et al., 2008; Lin and Kuo, 2010; Surdeanu et al., 2011; Ture and Boschee, 2016). They can also map terms across languages using Wikipedia links or BabelNet (Bouma et al., 2008; Pouran Ben Veyseh, 2016). However, *adversarial training* has not been tried in that setting.

q: *give tips? did you do with it; if the answer is yes, then what the magnitude of what you avoid it? In our country, we leave a 15-20 percent.*

q'_1 Tipping in Qatar. Is Tipping customary in Qatar ? What is considered "reasonable" amount to tip : 1. The guy that pushes the shopping trolley for you 2. The person that washes your car 3. The tea boy that makes coffee for you in the office 4. The waiters at 5-star restaurants 5. The petrol pump attendants etc
Relevant

q'_2 Tipping Beauty Salon. What do you think how much i should tip the stuff in a beauty salon for manicure/pedicure; massage or haircut?? Any idea what is required in Qatar?
Relevant

...

q'_9 Business Meeting? Guys; I'm just enquiring about what one should wear to business meetings in Doha? Are there certain things a man should or shouldn't wear (Serious replys only - not like A man shouldn't wear a dress)!!!! Thanks - Gino
Irrelevant

q'_{10} what to do? I see this man every morning; cleaning the road. I want to give him some money(not any big amount)but I feel odd to offer money to a person who is not asking for it. I am confused; I kept the money handy in the car.... because of the traffic the car moves very slowly in that area; I can give it to him easily..but am not able to do it for the past 4 days; and I feel so bad about it. If I see him tomorrow; What to do?
Irrelevant

Figure 1: An input question and some of the potentially relevant questions retrieved for it.

3 Adversarial Training for Cross-Language Problems

We demonstrate our approach for cross-language representation learning with adversarial training on a cross-lingual extension of the *question–question similarity reranking* subtask of SemEval-2016 Task 3 on community Question Answering.

An example for the monolingual task is shown in Figure 1. We can see an original English input question q and a list of several potentially similar questions q'_i from the Qatar Living[1] forum, retrieved by a search engine. The original question (also referred to as a new question) asks about how to tip in Qatar. Question q'_1 is relevant with respect to it as it asks the same thing, and so is q'_2, which asks how much one should tip in a specific situation. However, q'_9 and q'_{10} are irrelevant: the former asks about what to wear at business meetings, and the latter asks about how to tip a kind of person who does not normally receive tips.

[1] http://www.qatarliving.com/forum

In our case, the input question q is in a different language (Arabic) than the language of the retrieved questions (English). The goal is to rerank a set of K retrieved questions $\{q'_k\}_{k=1}^{K}$ written in a source language (e.g., English) according to their similarity with respect to an input user question q that comes in another (target) language, e.g., Arabic. For simplicity, henceforth we will use Arabic as target and English as source. However, in principle, our method generalizes to any source-target language pair.

3.1 Unsupervised Language Adaptation

We approach the problem as a classification task, where given a question pair (q, q'), the goal is to decide whether the retrieved question q' is *similar* (i.e., relevant) to q or not. Let $c \in \{0, 1\}$ denote the class label: 1 for similar, and 0 for not similar. We use the posterior probability $p(c = 1|q, q', \theta)$ as a score for ranking all retrieved questions by similarity, where θ are the model parameters.

More formally, let $\mathcal{R}_n = \{q'_{n,k}\}_{k=1}^{K}$ denote the set of K retrieved questions for a new question q_n. Note that the questions in $\mathcal{R}_n$ are always in English. We consider a training scenario where we have labeled examples $\mathcal{D}_S = \{q_n, q'_{n,k}, c_{n,k}\}_{n=1}^{N}$ for English q_n, but we only have unlabeled examples $\mathcal{D}_T = \{q_n, q'_{n,k}\}_{n=N+1}^{M}$ for Arabic q_n, with $c_{n,k}$ denoting the class label for the pair $(q_n, q'_{n,k})$. We want to train a cross-language model that can classify any test example $\{q_n, q'_{n,k}\}$, where q_n is in Arabic. This scenario is of practical importance, e.g., when an Arabic speaker wants to query the system in Arabic, and the database of related information is only in English. Here, we adapt the idea for adversarial training for domain adaptation as proposed by Ganin et al. (2016).

Figure 2 shows the architecture of our cross-language adversarial neural network (CLANN) model. The input to the network is a pair (q, q'), which is first mapped to fixed-length vectors $(\mathbf{z}_q, \mathbf{z}_{q'})$. To generate these word embeddings, one can use existing tools such as *word2vec* (Mikolov et al., 2013) and monolingual data from the respective languages. Alternatively, one can use cross-language word embeddings, e.g., trained using the *bivec* model (Luong et al., 2015). The latter can yield better initialization, which could be potentially crucial when the labeled data is too small to train the input representations with the end-to-end system.

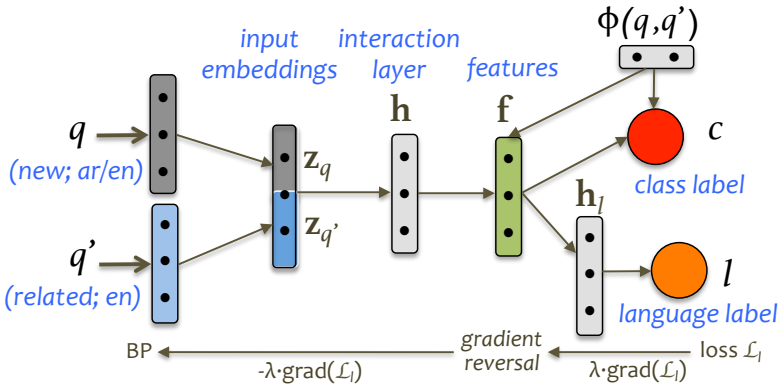

Figure 2: Architecture of CLANN for the question to question similarity problem in cQA.

The network then models the interactions between the input embeddings by passing them through two non-linear hidden layers, $\mathbf{h}$ and $\mathbf{f}$. Additionally, the network considers *pairwise* features $\phi(q, q')$ that go directly to the output layer, and also through the second hidden layer.

The following equations describe the transformations through the hidden layers:

$$\mathbf{h} = g(U[\mathbf{z}_q; \mathbf{z}_{q'}]) \quad (1)$$

$$\mathbf{f} = g(V[\mathbf{h}; \phi(q, q')]) \quad (2)$$

where $[.;.]$ denotes concatenation of two column vectors, U and V are the weight matrices in the first and in the second hidden layer, and g is a nonlinear activation function; we use rectified linear units or ReLU (Nair and Hinton, 2010).

The pairwise features $\phi(q, q')$ encode different types of similarity between q and q', and task-specific properties that we describe later in Section 4. In our earlier work (Martino et al., 2017), we found it beneficial to use them directly to the output layer as well as through a hidden-layer transformation. The non-linear transformation allows us to learn high-level abstract features from the raw similarity measures, while the adversarial training, as we describe below, will make these abstract features language-invariant.

The output layer computes a sigmoid:

$$\hat{c}_\theta = p(c = 1|\mathbf{f}, \mathbf{w}) = \text{sigm}(\mathbf{w}^T[\mathbf{f}; \phi(q, q')]) \quad (3)$$

where $\mathbf{w}$ are the output layer weights.

We train the network by minimizing the negative log-probability of the gold labels:

$$\mathcal{L}_c(\theta) = -c \log \hat{c}_\theta - (1 - c) \log (1 - \hat{c}_\theta) \quad (4)$$

The network described so far learns the abstract features through multiple hidden layers that are discriminative for the classification task, i.e., *similar* vs. *non-similar*. However, our goal is also to make these features invariant across languages. To this end, we put a language discriminator, another neural network that takes the internal representation of the network $\mathbf{f}$ (see Equation 2) as input, and tries to discriminate between *English* and *Arabic* inputs — in our case, whether the input comes from $\mathcal{D}_S$ or from $\mathcal{D}_T$.

The language discriminator is again defined by a sigmoid function:

$$\hat{l}_\omega = p(l = 1|\mathbf{f}, \omega) = \text{sigm}(\mathbf{w}_l^T \mathbf{h}_l) \quad (5)$$

where $l \in \{0, 1\}$ denotes the language of q (1 for English, and 0 for Arabic), $\mathbf{w}_l$ are the final layer weights of the discriminator, and $\mathbf{h}_l = g(U_l \mathbf{f})$ defines the hidden layer of the discriminator with U_l being the layer weights and g being the ReLU activations.

We use the negative log-probability as the discrimination loss:

$$\mathcal{L}_l(\omega) = -l \log \hat{l}_\omega - (1 - l) \log \left(1 - \hat{l}_\omega\right) \quad (6)$$

The overall training objective of the composite model can be written as follows:

$$\mathcal{L}(\theta, \omega) = \sum_{n=1}^{N} \mathcal{L}_c^n(\theta) - \lambda \Big[\sum_{n=1}^{N} \mathcal{L}_l^n(\omega) + \sum_{n=N+1}^{M} \mathcal{L}_l^n(\omega) \Big] \quad (7)$$

where $\theta = \{U, V, \mathbf{w}\}$, $\omega = \{U, V, \mathbf{w}, U_l, \mathbf{w}_l\}$, and the hyper-parameter λ controls the relative strength of the two networks.

In training, we look for parameter values that satisfy a min-max optimization criterion as follows:

$$\theta^* = \underset{U,V,\mathbf{w}}{\operatorname{argmin}} \max_{U_l,\mathbf{w}_l} \mathcal{L}(U, V, \mathbf{w}, U_l, \mathbf{w}_l) \quad (8)$$

which involves a maximization (gradient ascent) with respect to $\{U_l, \mathbf{w}_l\}$ and a minimization (gradient descent) with respect to $\{U, V, \mathbf{w}\}$. Note that maximizing $\mathcal{L}(U, V, \mathbf{w}, U_l, \mathbf{w}_l)$ with respect to $\{U_l, \mathbf{w}_l\}$ is equivalent to minimizing the discriminator loss $\mathcal{L}_l(\omega)$ in Equation (6), which aims to improve the discrimination accuracy. In other words, when put together, the updates of the shared parameters $\{U, V, \mathbf{w}\}$ for the two classifiers work adversarially with respect to each other.

In our gradient descent training, the above min-max optimization is performed by reversing the gradients of the language discrimination loss $\mathcal{L}_l(\omega)$, when they are backpropagated to the shared layers. As shown in Figure 2, the gradient reversal is applied to layer $\mathbf{f}$ and also to the layers that come before it.

Our optimization setup is related to the training method of Generative Adversarial Networks or GANs (Goodfellow et al., 2014), where the goal is to build deep generative models that can generate realistic images. The discriminator in GANs tries to distinguish real images from model-generated images, and thus the training attempts to minimize the discrepancy between the two image distributions, i.e., *empirical* as in the training data vs. *model-based* as produced by the generator. When backpropagating to the generator network, they consider a slight variation of the reverse gradients with respect to the discriminator loss. In particular, if ρ is the discriminator probability, instead of reversing the gradients of $\log(1-\rho)$, they use the gradients of $\log \rho$. Reversing the gradient is a different way to achieve the same goal.

Training. Algorithm 1 shows pseudocode for the algorithm we use to train our model, which is based on stochastic gradient descent (SDG). We first initialize the model parameters by using samples from glorot-uniform distribution (Glorot and Bengio, 2010). We then form minibatches of size b by randomly sampling $b/2$ labeled examples from $\mathcal{D}_S$ and $b/2$ unlabeled examples from $\mathcal{D}_T$. For the labeled instances, both $\mathcal{L}_c(\theta)$ and $\mathcal{L}_l(\omega)$ losses are active, while only the $\mathcal{L}_l(\omega)$ loss is active for the unlabeled instances.

Algorithm 1: Model Training with SGD

Input : data $\mathcal{D}_S$, $\mathcal{D}_T$, batch size b
Output : learned model parameters $\{U, V, \mathbf{w}, U_l, \mathbf{w}_l\}$

1. Initialize model parameters;
2. **repeat**
   - (a) Randomly sample $\frac{b}{2}$ labeled examples from $\mathcal{D}_S$
   - (b) Randomly Sample $\frac{b}{2}$ unlabeled examples from $\mathcal{D}_T$
   - (c) Compute $\mathcal{L}_c(\theta)$ and $\mathcal{L}_l(\omega)$
   - (d) Take a gradient step for $\frac{2}{b}\nabla_\theta \mathcal{L}_c(\theta)$
   - (e) Take a gradient step for $\frac{2\lambda}{b}\nabla_{U_l,\mathbf{w}_l}\mathcal{L}_l(\omega)$
   - `// Gradient reversal`
   - (f) Take a gradient step for $-\frac{2\lambda}{b}\nabla_\theta \mathcal{L}_l(\omega)$

until *convergence*;

As mentioned above, the main challenge in adversarial training is to balance the two components of the network. If one component becomes smarter, its loss to the shared layer becomes useless, and the training fails to converge (Arjovsky et al., 2017). Equivalently, if one component gets weaker, its loss overwhelms that of the other, causing training to fail. In our experiments, the language discriminator was weaker. This could be due to the use of *cross-language* word embeddings to generate input embedding representations for q and q'. To balance the two components, we would want the error signals from the discriminator to be fairly weak initially, with full power unleashed only as the classification errors start to dominate. We follow the weighting schedule proposed by Ganin et al. (2016, p. 21), who initialize λ to 0, and then change it gradually to 1 as training progresses. I.e., we start training the task classifier first, and we gradually add the discriminator's loss.

3.2 Semi-supervised Extension

Above we considered an unsupervised adaptation scenario, where we did not have any labeled instance for the target language, i.e., when the new question q_n is in Arabic. However, our method can be easily generalized to a semi-supervised setting, where we have access to some labeled instances in the target language, $\mathcal{D}_{T^*} = \{q_n, \mathcal{R}_n, c_n\}_{n=M+1}^{L}$. In this case, each minibatch during training is formed by labeled instances from both $\mathcal{D}_S$ and $\mathcal{D}_{T^*}$, and unlabeled instances from $\mathcal{D}_T$.

System	Input	Discrim.	Target	Hyperparam. (b, d, $\mathbf{h}$, $\mathbf{f}$, l_2)	MAP	MRR	AvgRec
FNN	en	–	ar	8, 0.2, 10, 100, 0.03	75.28	84.26	89.48
CLANN	en	en vs. ar'	ar	8, 0.2, 15, 100, 0.02	**76.64**	**84.52**	**90.92**
FNN	ar	–	en	8, 0.4, 20, 125, 0.03	75.32	84.17	89.26
CLANN	ar	ar vs. en'	en	8, 0.4, 15, 75, 0.02	**76.70**	**84.52**	**90.61**

Table 1: Performance on the test set for our cross-language systems, with and without adversarial adaptation (CLANN and FNN, respectively), and for both language directions (en-ar and ar-en). The prime notation under the *Discrim.* column represents using a counterpart from the unlabeled data.

4 Experimental Setting

In this section, we describe the datasets we used, the generation of the input embeddings, the nature of the pairwise features, and the general training setup of our model.

4.1 Datasets

SemEval-2016 Task 3 (Nakov et al., 2016b), provides 267 input questions for training, 50 for development, and 70 for testing, and ten times as many potentially related questions retrieved by an IR engine for each input question: 2,670, 500, and 700, respectively. Based on this data, we simulated a *cross-language setup* for question-question similarity reranking. We first got the 387 original train+dev+test questions translated into Arabic by professional translators. Then, we used these Arabic questions as an input with the goal to rerank the ten related English questions. As an example, this is the Arabic translation of the original English question from Figure 1:

هل تعطون الاكراميات؟ ماذا تفعلون بهذا الشأن؛
اذا كانت الاجابة نعم ، ما هو قوة ما تتجنبونه؟
في بلادنا ، نترك من ١٥ إلى ٢٠ بالمئة.

We further collected 221 additional original questions and 1,863 related questions as unlabeled data, and we got the 221 English questions translated to Arabic.[2]

4.2 Cross-language Embeddings

We used the TED (Abdelali et al., 2014) and the OPUS parallel Arabic–English bi-texts (Tiedemann, 2012) to extract a bilingual dictionary, and to learn cross-language embeddings. We chose these bi-texts as they are conversational (TED talks and movie subtitles, respectively), and thus informal, which is close to the style of our community question answering forum.

We trained Arabic-English cross-language word embeddings from the concatenation of these bi-texts using *bivec* (Luong et al., 2015), a bilingual extension of *word2vec*, which has achieved excellent results on semantic tasks close to ours (Upadhyay et al., 2016). In particular, we trained 200-dimensional vectors using the parameters described in (Upadhyay et al., 2016), with a context window of size 5 and iterating for 5 epochs. We then compute the representation for a question by averaging the embedding vectors of the words it contains. Using these cross-language embeddings allows us to compare directly representations of an Arabic or an English input question q to English potentially related questions q'_i.

4.3 Pairwise Features

In addition to the embeddings, we also used some pairwise features that model the similarity or some other relation between the input question and the potentially related questions.[3] These features were proposed in the previous literature for the question–question similarity problem, and they are necessary to obtain state-of-the-art results.

In particular, we calculated the similarity between the two questions using machine translation evaluation metrics, as suggested in (Guzmán et al., 2016). In particular, we used BLEU (Papineni et al., 2002); NIST (Doddington, 2002); TER v0.7.25 (Snover et al., 2006); METEOR v1.4 (Lavie and Denkowski, 2009) with paraphrases; Unigram PRECISION; Unigram RECALL. We also used features that model various components of BLEU, as proposed in (Guzmán et al., 2015): n-gram precisions, n-gram matches, total number of n-grams (n=1,2,3,4), hypothesis and reference length, length ratio, and brevity penalty.

[2]Our cross-language dataset and code are available at https://github.com/qcri/CLANN

[3]This required translating the Arabic input question to English. For this, we used an in-house Arabic–English phrase-based statistical machine translation system, trained on the TED and on the OPUS bi-texts; for language modeling, it also used the English Gigaword corpus.

We further used as features the cosine similarity between question embeddings. In particular, we used (*i*) 300-dimensional pre-trained Google News embeddings from (Mikolov et al., 2013), (*ii*) 100-dimensional embeddings trained on the entire Qatar Living forum (Mihaylov and Nakov, 2016), and (*iii*) 25-dimensional Stanford neural parser embeddings (Socher et al., 2013). The latter are produced by the parser internally, as a by-product.

Furthermore, we computed various task-specific features, most of them introduced in the 2015 edition of the SemEval task by (Nicosia et al., 2015; Joty et al., 2015). This includes some question-level features: (1) number of URLs/images/emails/phone numbers; (2) number of tokens/sentences; (3) average number of tokens; (4) type/token ratio; (5) number of nouns/verbs/adjectives/adverbs/ pronouns; (6) number of positive/negative smileys; (7) number of single/double/ triple exclamation/interrogation symbols; (8) number of interrogative sentences (based on parsing); (9) number of words that are not in WORD2VEC's Google News vocabulary. Also, some question-question pair features: (10) count ratio in terms of sentences/tokens/nouns/verbs/ adjectives/adverbs/pronouns; (11) count ratio of words that are not in WORD2VEC's Google News vocabulary. Finally, we also have one meta feature: (12) reciprocal rank of the related question in the list of related questions.

4.4 Model settings

We trained our CLANN model by optimizing the objective in Equation (7) using ADAM (Kingma and Ba, 2015) with default parameters. For this, we used up to 200 epochs. In order to avoid overfitting, we used dropout (Srivastava et al., 2014) of hidden units, l_2 regularization on weights, and *early stopping* by observing MAP on the development dataset —if MAP did not increase for 15 consecutive epochs, we exited with the best model recorded so far. We optimized the values of the hyper-parameters using grid search: for minibatch (b) size in $\{8, 12, 16\}$, for dropout (d) rate in $\{0.2, 0.3, 0.4, 0.5\}$, for $\mathbf{h}$ layer size in $\{10, 15, 20\}$, for $\mathbf{f}$ layer size in $\{75, 100, 125\}$, and for l_2 strength in $\{0.01, 0.02, 0.03\}$. The fifth column in Table 1 shows the optimal hyper-parameter setting for the different models. Finally, we used the best model as found on the development dataset for the final evaluation on the test dataset.

	System	MAP	MRR	AvgRec
Monolingual (English) from SemEval-2016				
1.	IR rank	*74.75*	*83.79*	*88.30*
2.	UH-PRHLT (1st)	76.70	83.02	90.31
3.	ConvKN (2nd)	76.02	84.64	90.70
Cross-language (Arabic into English)				
4.	CLANN	76.70	84.52	90.61

Table 2: Comparison of our cross-language approach (CLANN) to the best results at SemEval-2016 Task 3, subtask B.

5 Evaluation Results

Below we present the experimental results for the unsupervised and semi-supervised language adaptation settings. We compare our cross-language adversarial network (CLANN) to a feed forward neural network (FNN) that has no adversarial part.

5.1 Unsupervised Adaptation Experiments

Table 1 shows the main results for our cross-language adaptation experiments. Rows 1-2 present the results when the target language is Arabic and the system is trained with English input. Rows 3-4 show the reverse case, i.e., adaptation into English when training on Arabic. *FNN* stands for *feed-forward neural network*, and it is the upper layer in Figure 2, excluding the language discriminator. *CLANN* is the full *cross-language adversarial neural network*, training the discriminator with English inputs paired with random Arabic related questions from the unlabeled dataset. We show three ranking-oriented evaluation measures that are standard in the field of Information Retrieval: mean average precision (MAP), mean reciprocal rank (MRR), and average recall (AvgRec). We computed them using the official scorer from SemEval-2016 Task 3.[4] Similarly to that task, we consider Mean Average Precision (MAP) as the main evaluation metric. The table also presents, for reproducibility, the values of the neural network hyper-parameters after tuning (in the fifth column).

We can see that the MAP score for FNN with Arabic target is 75.28. When doing the adversarial adaptation with the unlabeled Arabic examples (CLANN), the MAP score is boosted to 76.64 (+1.36 points). Going in the reverse direction, with English as the target, yields very comparable results: MAP goes from 75.32 to 76.70 (+1.38).

[4] http://alt.qcri.org/semeval2016/task3/

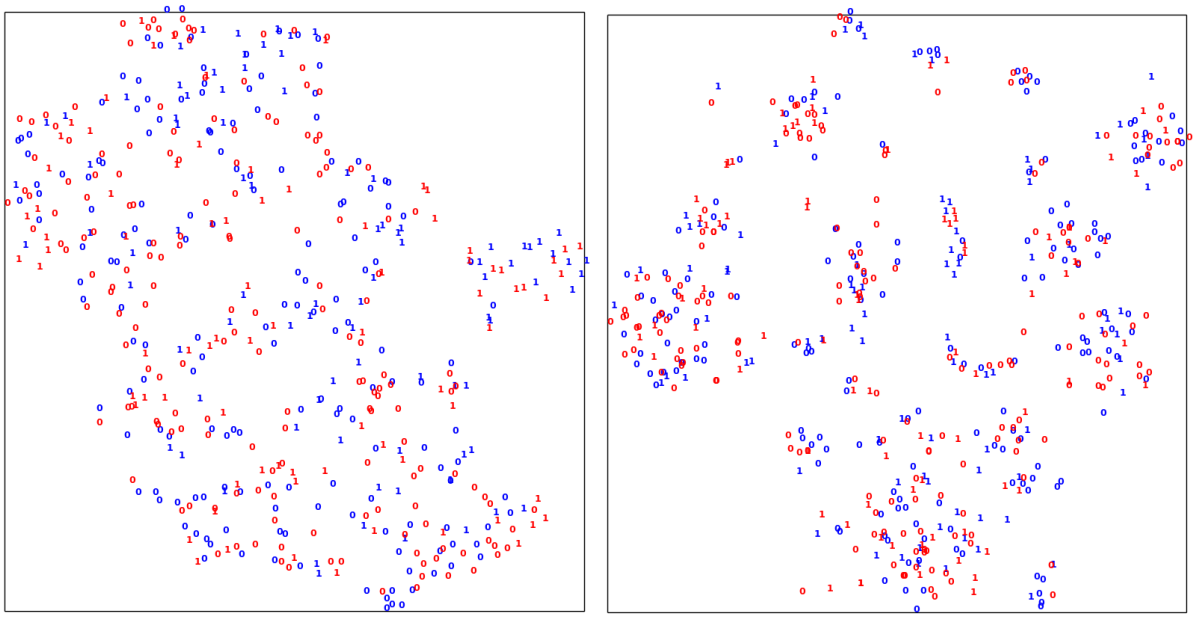

Figure 3: Scatter plots showing Arabic and English test examples, after training the adversarial network. Arabic is shown in blue, and English is in red. 0-1 are the class labels. Left: ar→en, right: en→ar.

To put these results into perspective, Table 2 shows the results for the top-2 best-performing systems from SemEval-2016 Task 3, which used a monolingual English setting. We can see that our FNN approach based on cross-language input embeddings is already not far from the best systems. Yet, when we consider the full adversarial network, in any of the two directions, we get performance that is on par with the best, in all metrics.

We conclude that the adversarial component in the network does the expected job, and improves the performance by focusing the language-independent features in the representation layer. The scatter plots in Figure 3 are computed by projecting the representation layer vectors of the first 500 test examples into two dimensions using t-SNE visualization (van der Maaten and Hinton, 2008). The first 250 are taken with Arabic input (blue), the second 250 are taken with English input (red). 0-1 are the class labels (similar vs. non-similar). The top plot corresponds to CLANN training with English and adapting with Arabic examples, while the second one covers the opposite direction. The plots look as expected. CLANN really mixes the blue and the red examples, as the adversarial part of the network pushes for learning shared abstract features that are language-insensitive. At the same time, the points form clusters with clear majorities of 0s or 1s, as the supervised part of the network learns how to classify them in these classes.

5.2 Semi-supervised Experiments

We now study the semi-supervised scenario when we also have some labeled data from the target language, i.e., where the original question q is in the target language. This can be relevant in practical situations, as sometimes we might be able to annotate some data in the target language. It is also an exploration of training with data in multiple languages all together.

To simulate this scenario, we split the training set in two halves. We train with one half as the source language, and we use the other half with the target language as extra supervised data. At the same time, we also use the unlabeled examples as before. We introduced the semi-supervised model in subsection 3.2, which is a straightforward adaptation of the CLANN model.

Table 3 shows the main results of our cross-language semi-supervised experiments. The table is split into two blocks by source and target language (en-ar or ar-en). We also use the same notation as in Table 1. The suffixes *-unsup* and *-semisup* indicate whether CLANN is trained in unsupervised mode (same as in Table 1) or in semi-supervised mode. The language discriminator in this setting is trained to discriminate between labeled source and labeled target examples, and labeled source and unlabeled target examples. This is indicated in the *Discrim.* column using asterisk and prime symbols, respectively.

System	Input	Discrim.	Target	Hyperparam. (b, d, $\mathbf{h}$, $\mathbf{f}$, l_2)	MAP	MRR	AvgRec
FNN	en	—	ar	8, 0.3, 10, 100, 0.03	74.69	83.79	88.16
CLANN-unsup	en	en vs. ar'	ar	12, 0.3, 15, 75, 0.02	75.93	84.15	89.63
CLANN-semisup	en+ar*	{en vs. ar* en vs. ar'	ar	8, 0.4, 15, 75, 0.02	76.65	84.52	90.84
FNN	ar	—	en	8, 0.2, 10, 75, 0.03	75.38	84.05	89.12
CLANN-unsup	ar	ar vs. en'	en	12, 0.2, 15, 75, 0.03	75.89	84.29	89.54
CLANN-semisup	ar+en*	{ar vs. en* ar vs. en'	en	8, 0.2, 10, 75, 0.03	76.63	84.52	90.82

Table 3: Semi-supervised experiments, when training on half of the training dataset, and evaluating on the full testing dataset. Shown is the performance of our cross-language models, with and without adversarial adaptation (i.e., using CLANN and FNN, respectively), using the unsupervised and the semi-supervised settings, and for both language directions: English–Arabic and Arabic–English. The prime notation in the *Discrim.* column represents choosing a counterpart for the discriminator from the unlabeled data. The asterisks stand for choosing an unpaired labeled example from the other half of the training dataset.

There are several interesting observations that we can make about Table 3. First, since here we are training with only 50% of the original training data, both FNN and CLANN-unsup yield lower results compared to before, i.e., compared to Table 1; this is to be expected. However, the unsupervised adaptation, i.e., using the CLANN-unsup model, still yields improvements over the FNN model by a sizable margin, according to all three evaluation measures. When we also train using the additional labeled examples in the target language, i.e., using the CLANN-semisup model, the results are boosted again to a final MAP score that is very similar to what we had obtained before with the full source-language training dataset. In the English into Arabic adaptation, the MAP score jumps from 74.69 to 76.65 (+1.96 points) when going from the FNN to the CLANN-semisup model, the MRR score goes from 83.79 to 84.52 (+0.73), and the AvgRec score is boosted from 88.16 to 90.84 (+2.68). The results in the opposite adaptation direction, i.e., from Arabic into English, follow a very similar pattern.

These results demonstrate the effectiveness and the flexibility of our general adversarial training framework within our CLANN architecture when applied to a cross-language setting for question-question similarity, taking advantage of the unlabeled examples in the target language (i.e., when using unsupervised adaptation) and also taking advantage of any labeled examples in the target language that we may have at our disposal (i.e., when using semi-supervised training with input examples in the two languages simultaneously).

6 Conclusion

We have studied the problem of cross-language adaptation for the task of question–question similarity reranking in community question answering, when the input question can be either in English or in Arabic with the objective to port a system trained on one input language to another input language given labeled data for the source language and only unlabeled data for the target language. We used a discriminative adversarial neural network, which we trained to learn task-specific representations directly. This is novel in a cross-language setting, and we have shown that it works quite well. The evaluation results have shown sizable improvements over a strong neural network model that uses simple projection with cross-language word embeddings.

In future work, we want to extend the present research in several directions. For example, we would like to start with monolingual word embeddings and to try to learn the shared cross-language representation directly as part of the end-to-end training of our neural network. We further plan to try LSTM and CNN for generating the initial representation of the input text (instead of simple averaging of word embeddings). We also want to experiment with more than two languages at a time. Another interesting research direction we want to explore is to try to adapt our general CLANN framework to other tasks, e.g., to answer ranking in community Question Answering (Joty et al., 2016; Nakov et al., 2016a) in a cross-language setting, as well as to cross-language representation learning for words and sentences.

Acknowledgment

This research was performed by the Arabic Language Technologies group at Qatar Computing Research Institute, HBKU, within the Interactive sYstems for Answer Search project (IYAS).

References

Ahmed Abdelali, Francisco Guzmán, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavik, Iceland, LREC '14, pages 1856–1862.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, CA, USA, SemEval '16, pages 497–511.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR* abs/1701.07875.

Gosse Bouma, Jori Mur, and Gertjan van Noord. 2008. Question answering with Joost at CLEF 2008. In *Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum: Evaluating Systems for Multilingual and Multimodal Information Access*. Aarhus, Denmark, CLEF '08, pages 257–260.

Yunbo Cao, Huizhong Duan, Chin-Yew Lin, Yong Yu, and Hsiao-Wuen Hon. 2008. Recommending questions using the MDL-based tree cut model. In *Proceedings of the 17th International Conference on World Wide Web*. Beijing, China, WWW '08, pages 81–90.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Antonio Uva, and Alessandro Moschitti. 2016. Learning to re-rank questions in community question answering using advanced features. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. Indianapolis, IN, USA, CIKM '16, pages 1997–2000.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*. San Diego, CA, USA, HLT '02, pages 138–145.

Cicero dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, ACL-IJCNLP '15, pages 694–699.

Abdessamad Echihabi and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, ACL '03, pages 16–23.

Pamela Forner, Anselmo Peñas, Eneko Agirre, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard F. E. Sutcliffe, and Erik F. Tjong Kim Sang. 2008. Overview of the CLEF 2008 Multilingual Question Answering Track. In *Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum: Evaluating Systems for Multilingual and Multimodal Information Access*. Aarhus, Denmark, CLEF '08, pages 262–295.

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

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Sardinia, Italy, volume 9 of *AISTATS '10*, pages 249–256.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems Conference 27*. Montréal, Canada, NIPS '14, pages 2672–2680.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. 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)*. Beijing, China, ACL-IJCNLP '15, pages 805–814.

Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2016. Machine translation evaluation meets community question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, ACL '16, pages 460–466.

Sven Hartrumpf, Ingo Glöckner, and Johannes Leveling. 2008. Efficient question answering with question decomposition and multiple answer streams. In *Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum: Evaluating Systems for Multilingual and Multimodal Information Access*. Aarhus,Denmark, CLEF '08, pages 421–428.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. Bremen, Germany, CIKM '05, pages 84–90.

Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global Thread-level Inference for Comment Classification in Community Question Answering. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, Lisbon, Portugal, pages 573–578.

Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2016. Joint learning with global inference for comment classification in community question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*. San Diego, CA, USA, NAACL-HLT '16.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*. San Diego, CA, USA, ICLR '15.

Alon Lavie and Michael Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation* 23(2–3):105–115.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi S. Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. San Diego, CA, USA, NAACL-HLT '16, pages 1279–1289.

Chuan-Jie Lin and Yu-Min Kuo. 2010. Description of the NTOU complex QA system. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. Tokyo, Japan, NTCIR '10, pages 47–54.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Denver, CO, USA, pages 151–159.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. 2016. Adversarial autoencoders. In *Proceedings of the International Conference on Learning Representations 2016*. San Juan, Puerto Rico, ICLR '16.

Giovanni Da San Martino, Salvatore Romeo, Alberto Barrón Cedeño, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2017. Cross-language question re-ranking. In *Proceedings of the 40th ACM SIGIR Conference on Research and Development in Information Retrieval*. Tokyo, Japan, SIGIR '17.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. 2017. Unrolled generative adversarial networks. In *5th International Conference on Learning Representations*. Toulon, France, ICLR '17.

Todor Mihaylov and Preslav Nakov. 2016. SemanticZ at SemEval-2016 Task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, CA, USA, SemEval '16, pages 879–886.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*. Atlanta, GA, USA, NAACL-HLT '13, pages 746–751.

Teruko Mitamura, Hideki Shima, Tetsuya Sakai, Noriko Kando, Tatsunori Mori, Koichi Takeda, Chin-Yew Song Ruihua Lin, Chuan-Jie Lin, and Cheng-Wei Lee. 2010. Overview of the NTCIR-8 ACLIA tasks: Advanced cross-lingual information access. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. Tokyo, Japan, NTCIR '10, pages 15–24.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research* 55(1):95–130.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*. Haifa, Israel, ICML '10, pages 807–814.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17.

Preslav Nakov, Lluís Màrquez, and Francisco Guzmán. 2016a. It takes three to tango: Triangulation approach to answer ranking in community question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX, USA, EMNLP '16, pages 1586–1597.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016b.

SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, CA, USA, SemEval '16, pages 525–545.

Massimo Nicosia, Simone Filice, Alberto Barrón-Cedeño, Iman Saleh, Hamdy Mubarak, Wei Gao, Preslav Nakov, Giovanni Da San Martino, Alessandro Moschitti, Kareem Darwish, Lluís Màrquez, Shafiq Joty, and Walid Magdy. 2015. QCRI: Answer selection for community question answering - experiments for Arabic and English. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, CO,USA, SemEval '15, pages 203–209.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meting of the Association for Computational Linguistics*. Philadelphia, PA, USA, ACL '02, pages 311–318.

Amir Pouran Ben Veyseh. 2016. Cross-lingual question answering using common semantic space. In *Proceedings of the 2016 Workshop on Graph-based Methods for Natural Language Processing*. San Diego, CA, USA, TextGraphs '16, pages 15–19.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the 4th International Conference on Learning Representations*. San Juan, Puerto Rico, ICLR '16.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, ACL '07, pages 464–471.

Salvatore Romeo, Giovanni Da San Martino, Alberto Barrón-Cedeño, Alessandro Moschitti, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Mitra Mohtarami, and James Glass. 2016. Neural attention for learning to rank questions in community question answering. In *Proceedings of the 26th International Conference on Computational Linguistics*. Osaka, Japan, COLING '16, pages 1734–1745.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Cambridge, MA, USA, AMTA '06, pages 223–231.

Ian Soboroff, Kira Griffitt, and Stephanie Strassel. 2016. The BOLT IR test collections of multilingual passage retrieval from discussion forums. In *Proceedings of the 39th International Conference on Research and Development in Information Retrieval*. Pisa, Italy, SIGIR '16, pages 713–716.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, ACL '13, pages 455–465.

Radu Soricut and Eric Brill. 2006. Automatic question answering using the web: Beyond the factoid. *Information Retrieval* 9(2):191–206.

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

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics* 37(2):351–383.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey, LREC '12, pages 2214–2218.

Ferhan Ture and Elizabeth Boschee. 2016. Learning to translate for multilingual question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX, USA, EMNLP '16, pages 573–584.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, ACL '16, pages 1661–1670.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 9:2579 – 2605.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based QA services. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. Boston, MA, USA, SIGIR '09, pages 187–194.

Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question retrieval with high quality answers in community question answering. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. Shanghai, China, CIKM '14, pages 371–380.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, OR, USA, ACL '11, pages 653–662.