Dual Constrained Question Embeddings with Relational Knowledge Bases for Simple Question Answering

Kaustubh Kulkarni and **Riku Togashi** and **Hideyuki Maeda** and **Sumio Fujita**
Yahoo Japan Corporation,
Tokyo, Japan
{kkulkarn,rtogashi,hidmaeda,sufujita}@yahoo-corp.jp

Abstract

Embedding based approaches are shown to be effective for solving simple Question Answering (QA) problems in recent works. The major drawback of current approaches is that they look only at the similarity (constraint) between a question and a head, relation pair. Due to the absence of tail (answer) in the questions, these models often require paraphrase datasets to obtain adequate embeddings. In this paper, we propose a dual constraint model which exploits the embeddings obtained by Trans* family of algorithms to solve the simple QA problem without using any additional resources such as paraphrase datasets. The results obtained prove that the embeddings learned using dual constraints are better than those with single constraint models having similar architecture.

1 Introduction

Recent progress in Knowledge Bases (KB) related technologies enable us to enhance an atomic fact repository of inter-entity relationship. For example KB completion aims to infer unknown entities in atomic facts, which is represented in the form of triplets (h, r, t) where h, r, t represent a head entity, relationship and a tail entity respectively, e.g. (*Barack Obama, Nationality, USA*) which corresponds to the factual knowledge that "the nationality of Barack Obama is USA". Freebase[1] and DBPedia[2] contain such atomic facts about entities in the real world. However, the real challenge for leveraging such knowledge in practical applications consists of mapping natural language questions to their corresponding entries in thus enhanced KBs.

Current embedding based QA models such as (Bordes et al., 2014a,b, 2015; Golub and He, 2016) are focusing on a sequential inference of predicting the pair of (h,r) from the given question (q), then inferring (t) corresponding to the predicted pair (h,r) using any KB completion models e.g. Trans* family models. This is a reasonable approach since such a type of questions contain information about both the head entity and the relation. However, once the first step of inference fails to match the correct (h,r) pair, it is hopeless for the second step to answer the correct entity. In order to avoid this problem, they use additional resources such as question paraphrases or entity aliases.

In this paper, we propose a completely different approach which uses a Trans* family based scoring function to predict the pair of (h,r) from q, and also maps q to t simultaneously. We learn embeddings for question words, entities and relations from the KB simultaneously bringing them into an euclidean space. Proposed dual constraint concurrent inference achieved better performance on a standard dataset than single constraint sequential inference methods without using any additional resources.

2 Related Work

Our work is inspired by the recent advances in solving simple QA problems using embedding approaches such as (Bordes et al., 2014a,b, 2015) which show that these approaches are very effective in mapping natural language questions to the corresponding triplet in a KB. They learn the embeddings for each question by a Bag Of Words (BOW) representation. (Jain, 2016) focuses on the positions of the question words and

---

[1] https://developers.google.com/freebase/
[2] http://wiki.dbpedia.org/

incorporate them into the model. Some models such as (Yih et al., 2015; Dai et al., 2016) focus on deep networks to encode question words and KB constituents. (Dai et al., 2016) have modeled the probability of predicting head and relation jointly, proposing two neural networks for learning the embeddings. (Golub and He, 2016) introduced attention mechanism for character level LSTM encoding of questions and entities for embedding question words and KB constituents. (Yin et al., 2016) proposed to use char-CNN to encode the entities and word-CNN with maxpooling to encode relations. (Yu et al., 2017) focused on the different granularity of relation representation. (Lukovnikov et al., 2017) have used rich entity information to get more powerful KB constituents encodings. Note that all these models focused on a single constraint sequential inference that uses similarities between q and (h,r) pairs to learn embeddings. Our method applies a dual constrained concurrent inference that uses similarities between q and t on top of (h,r) pairs while leveraging embeddings pre-trained by the Trans* family of algorithms.

Various Trans* family models such as **TransE** (Bordes et al., 2013),**TransH**(Wang et al., 2014),**TransR** (Lin et al., 2015) are proposed to learn low dimension embeddings of KB constituents using relations in KB as translations over entity embedding space.

3 Proposed Method

We propose an embedding based approach where both question words and KB constituents are mapped into a common low dimensional embedding space.

Single Constrained Sequential Inference: In the common euclidean space of question and KB constituents embeddings, assume d_1 to be euclidean distance between question embedding (**q**) and additive vector of head entity and relation (**h+r**). As shown in Figure 1[a] current QA models try to minimize d_1 so as to predict head entity and relation pair (h,r) from the question(q), consequently they have a single constraint such that corresponding **q** and **h+r** should be closer to each other. Assume d_2 to be an euclidean distance between tail entity embedding (**t**) and (**h+r**). Trans* family of algorithms try to minimize d_2 so as to predict tail entity (t) from a pair of (h,r). Current models minimize d_1 and d_2 in the two distinct
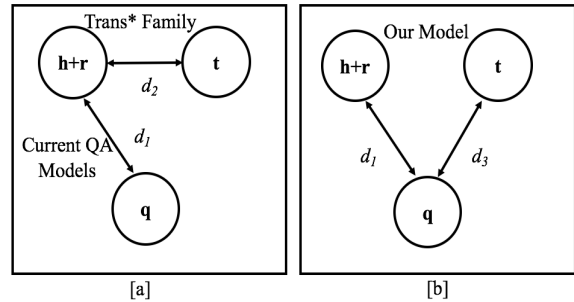

Figure 1: **[a]** Single constrained QA models minimize the distance (d_1) between Question Embeddings (q) and head entity, relation pair (h,r) and Trans* family of algorithms minimize distance (d_2) between tail entity (t) and (h,r). **[b]** Our model minimize the distance (d_3) between q and t along with d_1 thus applying the second constraint on q.

steps, indirectly bringing **q** closer to **t**.

Dual Constrained Concurrent Inference: A QA system is preferably able to directly retrieve an answer entity upon a submission of a question. The problem here is that a simple factoid question does not contain sufficient information of the answer entity by definition. Thus our model should learn question embeddings such that **q** should be closer to **t** as well as (**h+r**). Assume d_3 to be an euclidean distance between (**q**) and (**t**) in the same vector space as earlier model, as can be seen in Figure 1[b], our model minimizes $d_1 + d_3$ i.e. bringing **q** closer to both **h+r** and **t**. This is implemented as a dual constraint in objective function when learning **q**, which reduces the degree of freedom of **q**, resulting in a better euclidean space in comparison with single constrained models. Thanks to these constraints we do not need any additional resources such as question paraphrase datasets while achieving on a par performance with the current models.

3.1 TransR

TransR (Lin et al., 2015) is an algorithm to learn low dimensional vector representations of entities and relations in the Knowledge Base. TransR adopts a score function f_r to measure the credibility of a KB triplet *(h, r, t)* such that the score is low when *(h, r, t)* is likely to be true and high otherwise.

TransR represents entities and relations in distinct vector spaces i.e. *entity space* and *relation space*. For each triplet, let $\mathbf{h} \in \mathbb{R}^k, \mathbf{t} \in \mathbb{R}^k$ be an entity embedding of either head or tail respec-
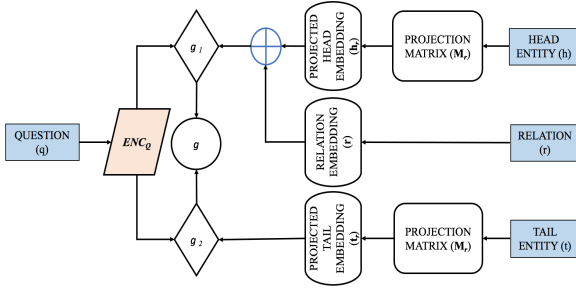
Figure 2: Overview of our learning diagram

tively and $\mathbf{r} \in \mathbb{R}^d$ be a relation embedding, where k and d are the dimensions of embeddings of entities and relations respectively. It also trains a relation specific projection Matrix $\mathbf{M}_r \in \mathbb{R}^{k \times d}$, which projects entities from the entity space to a corresponding relation space. With this projection matrix, projected vectors of entities are defined as,

$$\mathbf{h}_r = \mathbf{h}\mathbf{M}_r, \ \mathbf{t}_r = \mathbf{t}\mathbf{M}_r.$$

The score function is defined as:

$$f_r = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2.$$

Constraints are enforced on the norms of embeddings h, r, t and projection matrices, i.e. $\forall \ \mathbf{h}, \mathbf{r}, \mathbf{t}$ we have $\|\mathbf{h}\|_2 \leq 1, \|\mathbf{r}\|_2 \leq 1, \|\mathbf{t}\|_2 \leq 1, \|\mathbf{h}\mathbf{M}_r\|_2 \leq 1, \|\mathbf{t}\mathbf{M}_r\|_2 \leq 1.$

Margin based score function is defined as objective for training purpose which is as follows:

$$C = \sum_{(h, r, t) \in S} \sum_{(h', r, t') \in S'} max(0, f_r(h,t) + \gamma - f_r(h',t'))$$

where $max(x,y)$ returns the larger value between x and y, γ is the margin parameter, S is the set of correct triplets from dataset and S' is the set of corrupted triplets generated by using various negative sampling methods for training purpose.

3.2 Model

As shown in Figure 2, our model pre-trains the entity $(\mathbf{h},\mathbf{t})$ and relation $(\mathbf{r})$ embeddings with TransR, whereas encoding questions into $\mathbf{q} \in \mathbb{R}^d$ where d is the size of question embedding. Questions should be closer to the sum of (h) and (r) of the corresponding triplet in the KB, similarly the answer of the question should be closer to the vector of tail entity (t). We propose a score function $g(h, r, t, q)$ such that the score is low if (h, r, t) is the triplet corresponding to the question (q) and high otherwise. Scores of the question embedding are defined as:

1. This indicates how close a question is to the combination of head entity (h) and relation (r) embeddings in TransR relation space

$$g_1 = \|\mathbf{h}_r + \mathbf{r} - \mathbf{q}\|_2^2$$

2. This indicates how close a question is to the tail entity (t) embedding in a TransR relation space.

$$g_2 = \|\mathbf{t}_r - \mathbf{q}\|_2^2$$

Then the final score of the question is defined as:

$$g = g_1 + g_2.$$

Additional constraints are enforced on norms of embeddings such that $\|\mathbf{q}\|_2 \leq 1$. Due to a dual constraint mentioned above on the question embedding ($\mathbf{q}$), the degree of freedom is reduced considerably, which leads to fast training.

3.3 Training

Similar to previous studies involving embedding models (Bordes et al., 2014a,b, 2015), our model is trained with a ranking criterion. The objective of the learning is that the positive triplet should be closer to the natural language question than any other negative triplet by a certain margin γ in the embedding space. Thus we adopt a margin-based objective function for training purpose as follows:

$$L = \sum_{(h, r, t, q) \in S} \sum_{(h', r', t', q) \in S'} max(0, g_1(h, r, q) + \gamma - g_1(h',r', q)) + max(0, g_2(t,q) + \gamma - g_2(t',q))$$

where $max(x,y)$ and γ are same as defined earlier. S is the set of correct pairs of a triplet and a question from the dataset and S' is the set of pairs of a negative triplet and a question as S.

3.4 Negative Triplet generation

For generating negative triplets we use a method known as *candidates as negatives*, which is proposed by (Bordes et al., 2015). In this method, non-supported triplets are chosen randomly from the set of candidate triplets.

4 Experiments

4.1 Knowledge Base and Dataset

We use FB2M as our base KB which is an extract of the Freebase with about 2M entities and

5k relations. We use **SimpleQuestions**[3] dataset introduced by (Bordes et al., 2015) for training and testing purposes. This dataset consists of a total of 108,442 natural language questions in English written by human English speaking annotators each paired with a corresponding triplet from FB2M that provides the answer. Out of whole data 75,910 data points were used for Training, 10,845 for validation and 21,687 for testing purpose.

4.2 Experimental Setup

TransR embeddings TransR embeddings of size 64 initialized randomly with uniform distribution were pre-trained with Probabilistic Negative Sampling method proposed by (Kanojia et al., 2017).

Question Encoding Question is represented as sequence of words $(x_1, x_1, ..., x_{|q|})$. Low dimension vectors for each word in vocabulary are learnt and each word x_i is mapped to its vector. Word embedding size was set at 64 and initialized with random uniform distribution. We experimented with two methods to encode questions from individual question word embeddings:

Bag-of-Words (BOW) : It is a sum of individual word embeddings i.e.

$$\mathbf{q} = \sum_{i=1}^{|q|} x_i$$

Long Short Term Memory (LSTM)(Hochreiter and Schmidhuber, 1997): Each question is encoded using LSTM with dynamic RNN units with hidden layer size of 64 and forget bias as 1.0. Output of the last LSTM unit was taken as the question encoding.

We experimented with Batch size as 512, margin (γ) at 0.1 and Adam optimizer with learning rate of 0.001.

Candidate Pruning: Calculation of score for all triplets from the dataset is an memory and time wise prohibitive operation. Thus, at first we prune the facts to generate candidate facts similar to (Jain, 2016). Then only these candidate facts are scored by feeding them as input to our network. To generate these candidate facts, we match all possible n-grams of words of the question to Freebase entities and discard all n-grams (and their matched entities) that are a subsequence of another n-gram. All facts having one of the remaining entities as subject are added to candidate fact list. Facts with lowest score (g) out of candidates is retrieved as answer to the question. We evaluate our model based on

---

Setup	Path-level Accuracy(%)
Random Guess	4.9
Word Position(Jain, 2016)	59.7
Memory NW (Bordes et al., 2015)	62.7
Dual constrained BOW encoding	61.03
Dual constrained LSTM encoding	**64.05**

Table 1: Experimental Results on SimpleQuestion dataset for FB2M settings.

path-level accuracy in which prediction is correct if the head entity (h) and relation (r) of retrieved triplet are correct.

5 Results

The results of our experiments are shown in Table 1. We observe that our model gains 2-5% improvement in the path level accuracy than single constrained word level embeddings approaches by (Bordes et al., 2015; Jain, 2016) who have similar architecture to ours. Note that they use additional resources such as question paraphrase dataset and entity aliases while our model uses original dataset only. There are recent studies such as (Yin et al., 2016; Lukovnikov et al., 2017; Yu et al., 2017) which reported better accuracies on the same test set, by adopting either char-level CNNs or richer representations of entities/relations. Note that our dual constraint concurrent inference can be easily incorporated into such methods thus our method is complementary to their methods. We also report comparisons between different question encoding methods of our model. LSTM encoding outperforms BOW as it captures syntactic clues to map question onto the KB.

6 Conclusion and Future Work

In this work we show that Translation Embeddings learned using Trans* family of algorithms enable our model to learn the latent relationships between question and triplet using a unique score function. This results in a better performance in contrast with single constrained models as the essence of the triplet is inherently passed to the model in the form of embeddings. It also eliminates the need to use additional datasets to achieve good performance. The added dual constraint enforces the model to reduce the dual euclidean distance between question and triplet pairs, thereby generating adequate embeddings. Note that the dual constrained method can be extended to recent state of the art systems which use rich networks to obtain

better results.

In future, we hope to apply this method to richer embeddings obtained using deep networks. As shown in (Golub and He, 2016) character level encodings based models have been proven to be more precise compared to word level models for a simple QA task. We hope to extend our model to character level ones. Also the entity accuracy is comparatively lower than the relation accuracy which can be improved by using better entity linkers in questions.

References

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014a. Question answering with subgraph embeddings. In *EMNLP*, pages 615–620. ACL.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR*, abs/1506.02075.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26*, pages 2787–2795.

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014b. Open question answering with weakly supervised embedding models. In *ECML/PKDD (1)*, volume 8724 of *Lecture Notes in Computer Science*, pages 165–180. Springer.

Zihang Dai, Lei Li, and Wei Xu. 2016. CFO: conditional focused neural question answering with large-scale knowledge bases. *CoRR*, abs/1606.01994.

David Golub and Xiaodong He. 2016. Character-level question answering with attention. *CoRR*, abs/1604.00727.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Sarthak Jain. 2016. Question answering over knowledge base using factual memory networks. In *Proceedings of NAACL-HLT*, pages 109–115.

Vibhor Kanojia, Hideyuki Maeda, Riku Togashi, and Sumio Fujita. 2017. Enhancing knowledge graph embedding with probabilistic negative sampling. In *Proceedings of the 26th International Conference Companion on World Wide Web*, WWW '17 Companion, pages 801–802. International World Wide Web Conferences Steering Committee.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2181–2187. AAAI Press.

Denis Lukovnikov, Asja Fischer, Soeren Auer, and Jens Lehmann. 2017. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th international conference on World Wide Web*.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pages 1112–1119. AAAI Press.

Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. ACL Association for Computational Linguistics.

Wenpeng Yin, Mo Yu, Bing Xiang, Bowen Zhou, and Hinrich Schütze. 2016. Simple question answering by attentive convolutional neural network. *CoRR*, abs/1606.03391.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. *CoRR*, abs/1704.06194.