

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

Shafiq Joty, Preslav Nakov, Lluís Màrquez and Israa Jaradat
Qatar Computing Research Institute, HBKU

Problem Definition

Overall goal: learn cross-language representation of the input for the target task in a unified framework

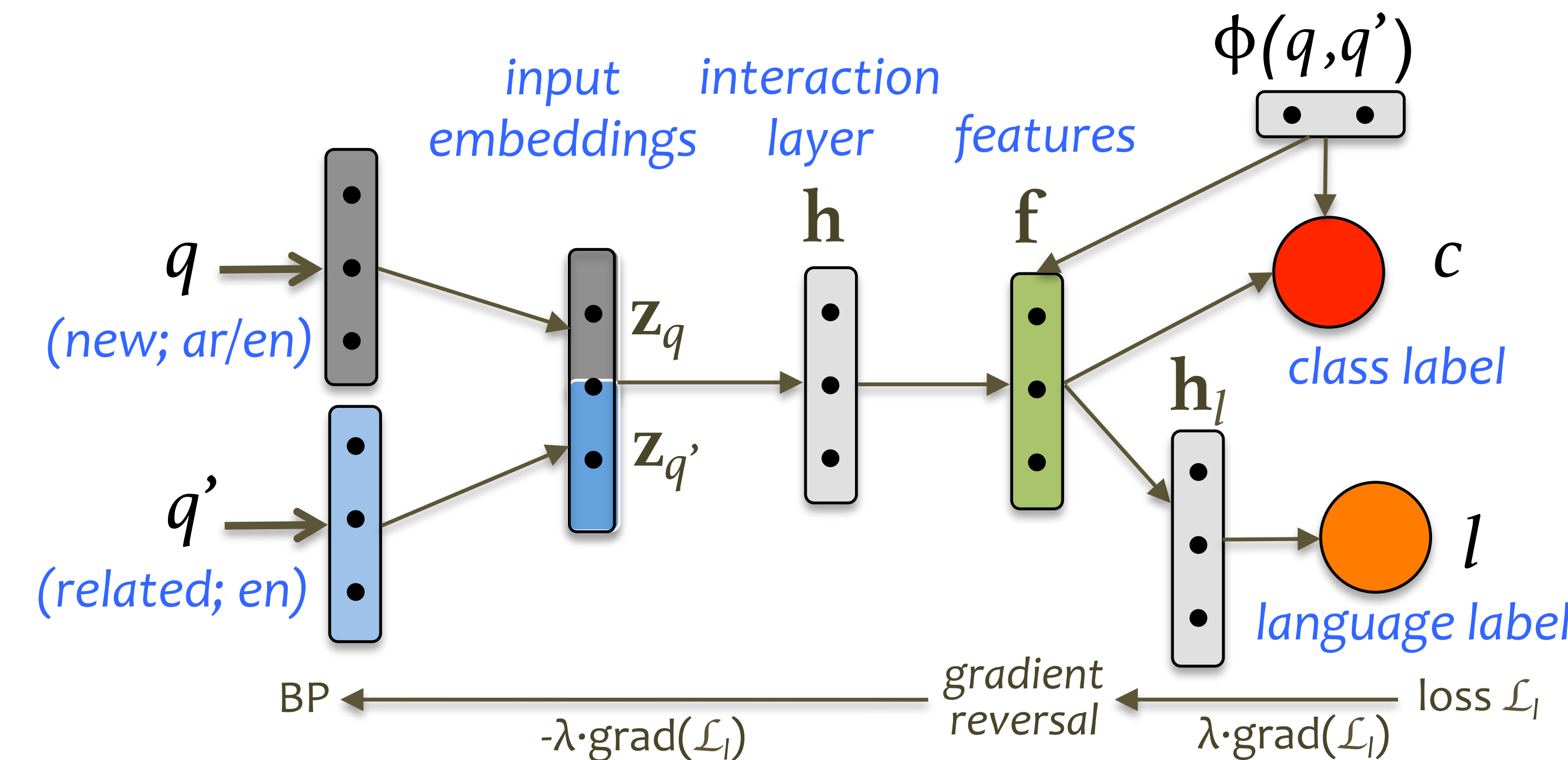
Cross-lingual question-question similarity scenario:

- **Input:**
 - ▷ a new user question q either in Arabic or in English
 - ▷ a set of potentially relevant existing questions $\{q'_k\}_{k=1}^K$, which are always in English
- **Task:** train a cross-lingual system to rerank $\{q'_k\}_{k=1}^K$ based on their similarity to q , where q is given in Arabic
- **Approach:**
 - ▷ train a binary classifier that decides whether q'_k is similar to q for a given pair of questions (q, q'_k)
 - ▷ use the posterior probability $p(c = 1|q, q'_k)$ for ranking
- **Training scenarios:**
 - ▷ *Unsupervised*: no class labels are given when q is in Arabic
 - ▷ *Semi-supervised*: some labeled examples available when q is in Arabic

Baseline Cross-lingual Model

- Use **cross-lingual embeddings** such as *bivec* (Luong et. al, 2015) to map q and q' to fixed-length vectors \mathbf{z}_q and $\mathbf{z}_{q'}$
 - ▷ yields better initialization
 - ▷ crucial when there is no enough labeled data to learn the input representations with end-to-end training
- Model interactions between \mathbf{z}_q and $\mathbf{z}_{q'}$:
 - ▷ $\mathbf{h} = g(U[\mathbf{z}_q; \mathbf{z}_{q'}])$
- Use pairwise features $\phi(q, q')$ to encode similarity directly:
 - ▷ $\mathbf{f} = g(V[\mathbf{h}; \phi(q, q')])$
 - ▷ $\phi(q, q')$ encode different similarity measures and task-specific features
 - ▷ A non-linear transformation allows us to learn high-level abstract features based on these pairwise features.
- The classification layer is defined by a sigmoid:
 - ▷ $\hat{c}_\theta = p(c = 1|\mathbf{f}, \mathbf{w}) = \text{sigm}(\mathbf{w}^T[\mathbf{f}; \phi(q, q')])$
- We optimize the log probability:
 - ▷ $\mathcal{L}_c(\theta) = -c \log \hat{c}_\theta - (1 - c) \log (1 - \hat{c}_\theta)$
- This network learns features 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.

Cross-Language Adversarial Neural Network (CLANN)



Adversarial Training

We put a **language discriminator**, another neural network that takes the internal representation of the network \mathbf{f} as input, and tries to discriminate between *English* and *Arabic* q .

- The discriminator is defined by another sigmoid: $\hat{l}_\omega = p(l = 1|\mathbf{f}, \omega) = \text{sigm}(\mathbf{w}_l^T \mathbf{h}_l)$
 - ▷ $\mathbf{h}_l = g(U_l \mathbf{f})$ defines the hidden layer of the discriminator
 - ▷ Discrimination loss: $\mathcal{L}_l(\omega) = -l \log \hat{l}_\omega - (1 - l) \log (1 - \hat{l}_\omega)$
- Overall training objective of the composite model:

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

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

- In training, we look for parameters that satisfy a min-max optimization criterion:

$$\theta^* = \underset{U, V, \mathbf{w}}{\text{argmin}} \max_{U_l, \mathbf{w}_l} \mathcal{L}(U, V, \mathbf{w}, U_l, \mathbf{w}_l) \quad (2)$$

The updates of the shared parameters $\{U, V, \mathbf{w}\}$ for the two classifiers is done in an adversarial way.

Features

- Cross-language embeddings trained with *bivec* on parallel corpora (TED talks and OPUS)
- Similarity-based pairwise features:
 - ▷ Machine Translation measures: BLEU, NIST, TER, METEOR, unigram PRECISION, unigram RECALL, and components of BLEU
 - ▷ Cosine similarity between questions: using Google and QatarLiving word embeddings, and Syntactic embeddings from the Stanford parser
 - ▷ Task-specific features (Joty et al., 2015).

Dataset

Based on the SemEval-2016 Task 3 dataset

- 387 original questions (276, 50, and 70 for training, development and test)
- For each original question 10 related questions to be ranked
- We translated the 387 original questions manually to Arabic.
- We further collected 221 original and 1,863 related questions (English; unlabeled). We manually translated the 221 questions to Arabic.

Results

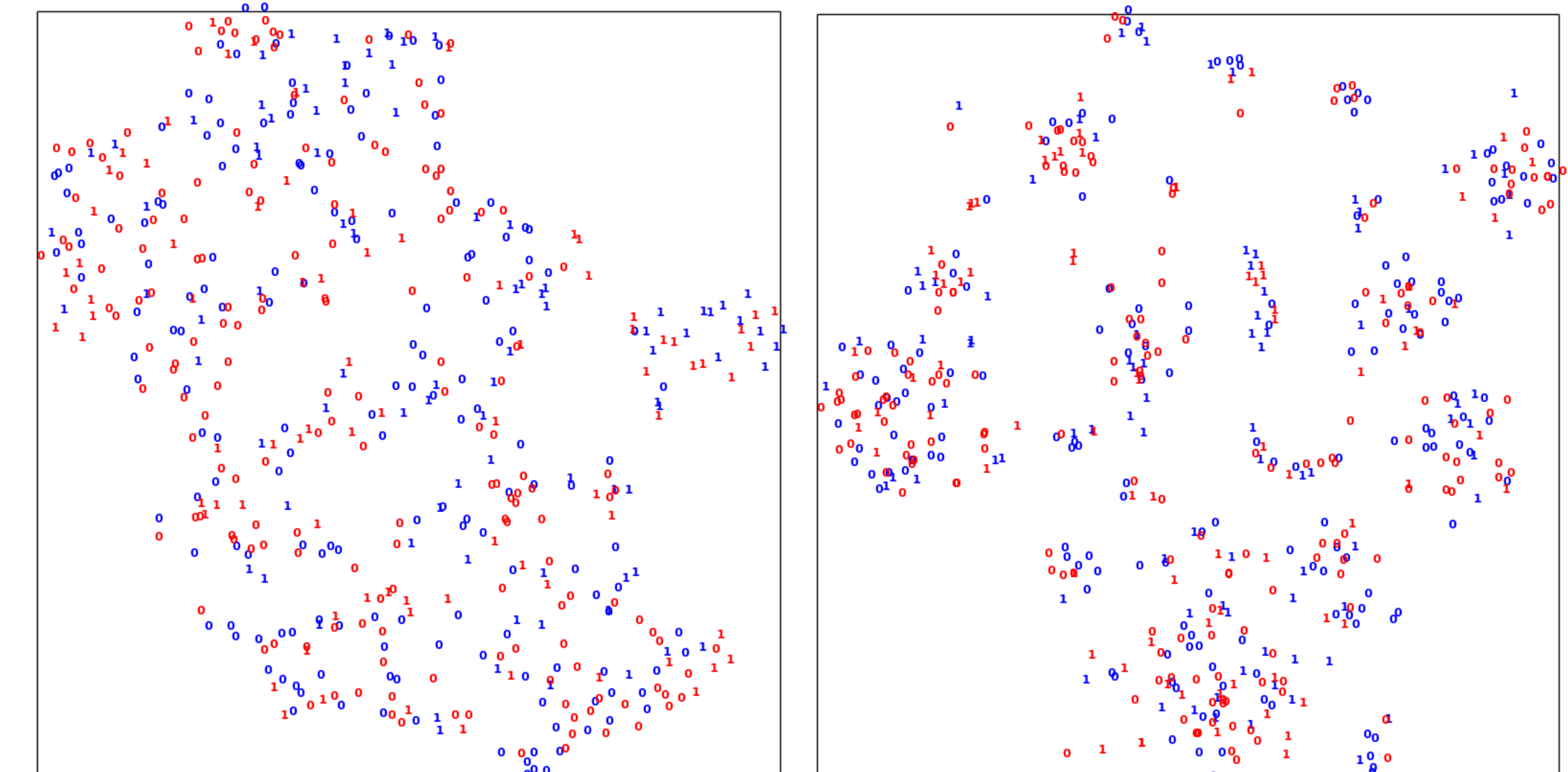
Unsupervised Adaptation

System		Discrim.	MAP	MRR	AvgRec
FNN	en→ar	—	75.28	84.26	89.48
CLANN	en→ar	en vs. ar'	76.64	84.52	90.92
FNN	ar→en	—	75.32	84.17	89.26
CLANN	ar→en	ar vs. en'	76.70	84.52	90.61

Semi-supervised Adaptation

System		Discrim.	MAP	MRR	AvgRec
FNN	en→ar	—	74.69	83.79	88.16
CLANN _{unsup}	en→ar	en vs. ar'	75.93	84.15	89.63
CLANN _{semisup}	en+ar*→ar	{en vs. ar* en vs. ar'}	76.65	84.52	90.84

Visualizing the Representation Layer



Arabic=blue, English=red. Class labels $\{0,1\}$. L: ar→en, R: en→ar

Conclusion

We have studied cross-language adaptation for question-question similarity in community question answering, in order to port a system trained on one input language to another input language. This is novel in a cross-language setting.

Future work

- Fine-tune the word embeddings for the cross-language task
- Try LSTM and CNN
- Experiment with more than two languages at a time
- Apply to other tasks