An Explorative Improvement of out-domain Question Answering using out-domain Fine-tuning and Word Emeddings

Phil Tinn

Abstract

This report describes a preliminary assessment of the effect of applying out-domain(OD) and domain-specific word embeddings and OD fine-tuning to a Question Answering(QA) system. The experiment used SQuAD, based on general questions and answers generated from Wikipedia articles, as its baseline model, and a biomedical-focused QA dataset, BioASQ, for OD QA evaluation. In building a model for OD QA, the performance contributions of non-domain specific (GloVe) and biomedical-focused (e.g. BioReddit, BioWordVec, etc.) word embeddings are compared. The revised model, a SQuAD(BioReddit) baseline fine-tuned with OD BioASQ data, achieved a score gain of +7.22 F1 and +3.0 EM over the baseline model in OD performance on BioASQ data. Meanwhile, the OD fine-tuning of a SQuAD (GloVe) baseline led to a gain of +5.36 F1 and +.27 EM. Comparing both OD fine-tuning cases, the use of in-domain (ID) word embeddings led to less gain in the EM score compared to the use of OD word embeddings. These preliminary results show favorable evidence for OD fine-tuning as a technique for improving the F1 score in OD QA task, and mild evidence for OD word-embedding in improving the EM score. They suggest promising improvement potentials in further applying Data Augmentation techniques to small but domain-specific datasets deployed for OD fine-tuning of models. And they highlight the need for a further in-depth comparison of tradeoffs in employing OD word embeddings for OD QA tasks.

1. Introduction

In recent years, Question Answering (QA) systems have become widely adopted across a variety of services that involve human's interfacing with information and knowledge (e.g. chatbot, virtual assistant, etc.). QA models, however, are often biased by their initial training dataset and therefore don't generalize well in tasks that involve a knowledge base in domains outside of that of the training dataset. Behind this problem are at least two barriers: first, out-of-domain QA datasets (e.g. BioASQ), even when available, can easily be significantly smaller than the in-domain QA dataset (e.g. SQuAD); secondly, Word Embeddings, the use of vector representations of words for capturing words' internal relationships and semantic properties, are becoming more tailored for biomedical NLP applications but also inherits the complexity of corpus localities across institutions and platforms [Wang et Al., 2018].

1.1 Hypothesis

Through the techniques of OD fine-tuning of the general QA model and applying OD word embeddings, the model's cross-domain performance is expected to increase. However, according to one key observation from a 2018 Mayo Clinic study, "word embeddings trained from biomedical domain corpora do not necessarily have better performance than those trained on other general domain corpora. That is, there might be no significant difference when word embeddings trained from an out-domain corpus are employed for a biomedical NLP application" [Wang et Al., 2018]. Taking this observation into account, to what extent would each technique affect the model's OD QA performance? To explore their respective influence on the model, an exploration was carried with several datasets (described below) for analysis.

2. Dataset

SQuAD: an open-domain reading comprehension QA dataset comprised of over 100,000 crowdsourced questions on a set of Wikipedia articles. This was used for the training of the baseline QA model.

BioASQ: a biomedical-focused reading comprehension QA dataset derived from PubMed articles and MeSH hierarchy for subject headings. This was divided into a training set and a development set, used respectively for OD fine-tuning on the baseline QA model and for OD QA evaluation.

GloVe 6B: a set of word embeddings created using an unsupervised learning algorithm GloVe on 6 billion tokens from Wikipedia articles with over 400k vocabularies. The evaluation covered 50, 100, 200, and 300 dimensions provided in the standard download pack.

BioReddit: word embeddings created from medical subreddits (~800,000 Reddit posts from over 60 medical-themed communities). The subreddits used, specifically, are *user-generated content* as opposed to biomedical literature like PubMed. For consistency with the format of open-domain GloVe word embedding, the BioReddit data downloaded for testing was in GloVe format in 200, 100, and 50 dimensions.

BioWordVec: a significantly larger set of word embeddings released by the National Center for Biotechnology Information (NCBI). It was created using the *subword embedding* model [Zhang et Al., 2019] and based on various collections of biomedical literature and domain knowledge in medical subject headings (MeSH). The data is presented in one "intrinsic" set for the semantic similarity between words and on "extrinsic" set for various downstream NLP tasks.

3. Experiment

To explore the relative effects of OD and ID word-embedding on cross-domain QA tasks, the study involved the following comparisons:

1. Given the range of dimensionality made available for GloVe (50, 100, 200, 300) and BioReddit (50, 100, 200), a quick comparison of the effect of embedding dimensionality on the OD QA task.
2. To observe the domain difference in embeddings on the training of the initial SQuAD baseline model, a comparison of baseline model built with GloVe embedding and one built with BioReddit embedding.
3. To observe the contribution of OD fine-tuning over the baseline model, additional training of SQuAD models from Part 2 using BioASQ data.
4. Speculatively, a third-round training of models from Part 3 using BioASQ data but with an alternative embedding (switch GloVe to BioReddit and vice versa).

4. Implementation

4.1 Model

The experiment made use of a prebuilt PyTorch implementation of a neural QA system based on DrQA, a system for reading comprehension applied to open-domain question-answer tasks [Chen et al, 2017]. The model uses bi-directional LSTM to encode input passages into a set of vectors and input questions into fixed vectors. The prediction of answers is computed as logits for both start pointer and end pointer to the span of token locations in the question passage. The experiment made use of the following parameters:

- RNN Cell: LSTM (Bidirectional)
- Batch Size: 64 / 128
- Learning Rate: 0.001 (with early stop of 3 epochs)
- Embedding Dimension: 50 / 100 / 200 / 300
- Vocabulary Size: 50,000 / 13,059
- Hidden Dimension: 256
- Max Context Length: 384
- Max Question Length: 64

- Dropout: 0.0
- Weight Decay: 0

The early fine-tuning attempts using the downloaded SQuAD pre-trained QA model resulted in a mismatch in shape: size mismatch for passage_rnn.weight_ih_l0: copying a param with shape torch.Size([512, 600]) from checkpoint, the shape in current model is torch.Size([1024, 600]). Using a locally trained baseline model would result in the shape of ([50002, 300]) Yet BioASQ, prior to its division into training and development sets, has a vocabulary size of 19889. Training based on 300 dimensions, for example, would lead to a ([19891, 300]) shape tensor. After splitting BioASQ data into training and development sets, the input vocabulary size was further reduced to 13,059. To meet the constraint of BioReddit embedding's 100 and 200 dimensions, the shape of the baseline model trained locally was ultimately adjusted to ([13061, 100]) and ([13061, 200]) respectively.

4.2 Dataset

Unlike the NewsQA and SQuAD datasets, the BioASQ dataset is smaller and comes without a development set. To operationalize the BioASQ dataset for OD fine-tuning of the baseline model, it was divided into a chunk of 769 QA samples for training and another chunk of 735 for development.

The .gz file of bioasq.json.gz was difficult to subdivide using Linux's split or gsplit command: they do not split the data where a QA sample ends but rather in the middle of a json object. After some failed attempts, the BioASQ was split manually for a clean division, and to give bioasq_train.jsonl and bioasq_dev.jsonl each a proper metadata header. A separate .jsonl loading path is created in def load_dataset(path) of utils.py.

Due to BioWordVec's large file size, it was shipped in binary format through the NCBI Github repository. The FastText library, imported through Gensim, was used to covert the BioWordVec's binary format to word2vec format to be consistent with that of GloVe word embeddings.

The training processes involving BioWordVec's word embeddings did not succeed, however. The RAM requirement needed to incorporate BioWordVec exceeded what Google Colab Pro+ and UT's computer lab server could support, both of which have a RAM limit of ~32GB. At best, the UT lab server trained up to only 300 batches with a batch size of 64 during the first epoch before it killed the process.

	50d	**100d**	**200d**	**300d**
	EM / F1			
Baseline (GloVe)	8.71 / 16.15	6.85 / 13.71	9.64 / 17.19	9.04 / 18.12
Baseline (BioReddit)	11.1 / 19.7	7.98 / 16.08	5.78 / 13.49	N/A
Baseline (BioWordVec)	N/A	N/A	Result Pending	N/A

Figure 1. Dimensionality Comparison of Word Embeddings in OD QA tasks

5. Observation

5.1 Embeddings Comparison

For building the baseline model on SQuAD, GloVe embedding outperformed BioReddit embedding across both 100d and 200d dimensions. The GloVe-trained baseline ([13061, 100]) gave an EM/F1 score of 11.7 / 19.08, towering over BioReddit's 5.85/13.75. But it's worth noting that when trained with a larger vocabulary size of ([50002, 50]), BioReddit was able to achieve an EM/F1 score of 11.1 / 19.7. Figure 2 consists of EM and F1 scores of the model trained with the GloVe word embeddings, and Figure 3 with BioRedding word embeddings.

Notably, while the fine-tuning of both models with different embeddings led to an observable gain in F1 score (+5.36 for GloVe at 100d, and +7.22 for BioReddit at 200d), there's almost no gain in EM score for the fine-tuned model trained with GloVe. On the other hand, the fine-tuned BioReddit model gained +3.95 and +3.0 EM in 100d and 200d. Given that the source used for BioReddit (300 million tokens and 780,000 vocabulary size) is 2-orders of magnitude smaller than the 6 billion tokens from Wikipedia used to generate GloVE, the modest lead the BioReddit had over GloVe can be considered remarkable.

	100d	**200d**
	EM / F1	
Baseline SQuAD (GloVe)	11.7 / 19.08	11.02 / 18.48
Baseline SQuAD + BioASQ (GloVe)	11.97 / 24.44 **+.27 / +5.36**	11.16 / 21.26 **+.14 / +2.78**
Baseline SQuAD + BioASQ (BioReddit)	6.26 / 14.8	N/A

Figure 2. GloVe-first Baseline with Fine-tuning: Using a baseline model trained on SQuAD with GloVe word embeddings and tested on BioASQ development set.

	100d	**200d**
	EM / F1	
Baseline SQuAD (BioReddit)	5.85 / 13.75	9.52 / 16.59
Baseline SQuAD + BioASQ (BioReddit)	9.8 / 19.55 **+3.95 /5.8**	12.52 / 23.81 **+3.0 / +7.22**
Baseline SQuAD + BioASQ (BioReddit) + BioASQ (GloVe)	5.58 / 13.25	N/A

Figure 3. BioReddit-first Baseline Fine-tuning: Using a baseline model trained on SQuAD with BioReddit word embeddings and tested on BioASQ development set.

When an additional round of retraining (in 100d) is applied to the already finetuned model using an alternative embedding (e.g. switch GloVe for BioReddit, and vice versa), in both cases, the OD performance drops significantly—back to near the original baseline level.

Given the difficulty in including BioWordVec word embeddings (~3GB) caused by this project's limited computational resource, the ~32GB RAM upper limit in Colab Pro+ ($50/month), only BioReddit and GloVe word embeddings were successfully assessed for their contribution to OD QA performance.

5.2 Parameter Selection

As one main part of the overall approach in this project to increase OD QA performance is to fine-tune a larger general model (e.g. based on SQuAD) with a domain-specific data set (e.g. BioASQ), inevitably, the baseline model used for retraining would encounter mismatches in their matrix shape due to its difference in vocabulary size with that of the dataset used for retraining.

Initially, the baseline model was trained with a vocabulary size of 50,000. However, during the retraining with the BioASQ data set, it had to downsize to match the BioASQ data's 19,889 vocabulary size. And with the further division of BioASQ data into training and development sets, the baseline model's vocabulary size was adjusted once more to 13,059 to match that of the BioASQ training set.

In reducing the vocabulary size, from 50000 to 19889 to 13061, we did not observe a significant reduction in validation score when using a SQuAD trained model to evaluate on SQuAD data. The EM/F1 scores from [(19891, 100]) to [(13061, 100]) were 48.46/61.11 and 48.11/60.64 respectively.

5.3 Dimensionality Evaluation

To explore the effect of dimensionality in the context of OD question answering, a brief

comparison was conducted across the performance of models trained on SQuAD with BioReddit embeddings in all three dimensions: 50d, 100d and 200d. (200-dimension embedding was the highest dimensional embedding made available through the BioReddit Github repository. The training result, shown in Figure 1, reveals an observable advantage of BioReddit embeddings in a lower dimension, with an EM/F1 score of 11.1 / 19.7, 7.98 / 16.08, and 5.78 / 13.49 for 50d, 100d and 200d.

This result runs contrary to the general perception that a small dimensionality lacks sufficient expression and that a very large dimensionality suffers from over-fitting [Yin et al., 2018]. On the other hand, models trained with GloVe embeddings exhibit a performance trajectory that increases along with each embedding's dimensionality, from 50d, 100d, 200d, to 300d, with an EM/F1 score of 8.71 / 16.15, 6.85 / 13.71, 9.64 / 17.19, and 9.04 / 18.12. The performance of GloVe embeddings in this OD QA conforms with 300 being the generally stable dimensionality since the introduction Word2Vec and GloVe models [Mikolove et al., 2013, Pennigton et al., 2014]. The difference shown in Figure 1 in the effect of dimensionality on GloVe embeddings and BioReddit embeddings presents an interesting question but is beyond the current scope of this project.

6. Conclusion & Future Steps

This paper provides an account of the investigation process that highlights the challenge in overcoming the gap between feature-rich training and feature-specific applications encountered in building QA systems that can generalize well for out-of-domain NLP tasks. This comparison of using OD fine-tuning and OD word embeddings in boosting the SQuAD model's OD QA tasks showed that model enrichment through fine-tuning with BioASQ's domain-focused data led to a more significant result. And the use of OD BioReddit word embeddings contributed principally to a modest gain in EM score.

While the benefit of OD fine-tuning is clear, more interestingly, the relative inconsequential advantage of biomedical word embedding over the general-purpose GloVe embedding further reveals the existence of unexplored nuances in using a biomedical-specific word embedding for driving the performance of downstream biomedical NLP applications—as highlighted in the prior comparison of word embeddings conducted by the Mayo Clinic for biomedical NLP processing [Wang et Al., 2018]. Nonetheless, given that BioReddit was created with just 300 million tokens, vis-a-vis GloVe's 6 billion, the superior cost-performance ratio of the domain-specific BioReddit word embedding is undeniable.

As the specific BioReddit word embedding deployed in this study was sourced from user-generated contents, instead of from across various professionally or academically-focused sources (e.g. PubMed) that hold the majority of knowledge produced in the biomedical field, it's necessary to further assess the potential usefulness of other biomedical word embeddings in OD QA. Therefore, it remains an open question how a significantly larger set of embeddings from BioWordVec, sourced through PubMed, will contribute to a better OD QA performance and shed more light on our understanding of the techniques and applications of OD word embeddings.

And lastly, as the most significant performance gain in OD QA was observed in OD fine-tuning using domain-focused but limited-size data from BioASQ, further work involving data augmentation to boost the size of limited OD QA samples may further lead to promising results.

References:

Marco Basaldella and Nigel Collier. 2019. BioReddit: Word Embeddings for User-Generated Biomedical NLP. In *Tenth International Workshop on Health Text Mining and Information Analysis*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Pages 3111–3119

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12 – 20

Zi Yin, Yuanyuan Shen. 2018. On the Dimensionality of Word Embedding. *Advances in Neural Information Processing Systems* 31

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu . 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data,* 6:52