Answer Passage Ranking Enhancement Using Shallow Linguistic Features

Bahadorreza Ofoghi[1] and Armita Zarnegar[2]

[1] School of Information Technology, Deakin University, Australia
b.ofoghi@deakin.edu.au
[2] Faculty of Science, Engineering, and Technology, Swinburne University of Technology, Australia
azarnegar@swin.edu.au

Abstract. Question Answering (QA) systems play an important role in decision support systems. Deep neural network-based passage rankers have recently been developed to more effectively rank likely answer-containing passages for QA purposes. These rankers utilize distributed word or sentence embeddings. Such distributed representations mostly carry semantic relatedness of text units in which explicit linguistic features are under-represented. In this paper, we take novel approaches to combine linguistic features (such as different part-of-speech measures) with distributed sentence representations of questions and passages. The QUASAR-T fact-seeking questions and short text passages were used in our experiments to show that while ensembling of deep relevance measures based on pure sentence embedding with linguistic features using several machine learning techniques fails to improve upon the passage ranking performance of our baseline neural network ranker, the concatenation of the same features within the network structure significantly improves the overall performance of passage ranking for QA.

Keywords: Question answering · Passage ranking · Deep learning · Shallow linguistic features.

1 Introduction

Natural language Question Answering (QA) systems have recently been utilized to support decision analysis and modeling (e.g., [7, 19]). Previous studies in the QA domain show that answer extraction is more effective from passage-level information compared with the analysis of full-text documents [12]. There is evidence of positive correlation between the effectiveness of QA and answer passage ranking [9]. For QA, both the general semantic relevance of a passage to the question and answer recall are of importance. For instance, given the question *"When did Google start?"*, the passage *"Google was launched by Larry Page and Sergey Brin, students at Stanford University"* will not be counted as an effective, answer-containing passage since it does not include the actual answer *1998*. Specificity of passages (i.e., containing answer candidates) in the QA domain

necessitates the utilization of explicit and shallow *linguistic* features, especially from within the passages to be considered in the process of passage ranking, those that are under-represented in the general semantics of sentences (see Figure 1).

	Semantically relevant passage	Specific answer-containing passage
Q1	The.det most.adv frequent.adj symptom.noun is.verb a.det stiff.adj jaw.noun, caused.v by.adp spasm.noun of.adp the.det muscle.noun that.adj closes.v the.det mouth.noun – accounting.v for.adp the.det disease.noun 's.part familiar.adj name.noun lockjaw.noun.	The.det first.adj sign.noun of.adp tetanus.noun is.v a.det tightening.noun of.adp the.det jaw.noun muscles.noun that.adj gives.v the.det disease.noun its.adj common.adj name.noun, lockjaw.noun.
Q2	He.prp composed.v many.adj operas.noun, but.conj his.adj greatest.adj triumph.noun was.v "I.noun Pagliacci.noun" for.adp which.adj he.prp wrote.v both.adj the.det libretto.noun and.conj the.det music.noun.	It.prp 's.v a.det setting.noun that.adj would.v have.v brought.v tears.noun of.adp joy.noun to.adp Ruggero.noun Leoncavallo.noun, the.det composer.noun and.conj librettist.noun who.noun gave.v Pagliacci.v life.noun in.adp 1892.num.
Q3	At.adp that.det time.noun, it.prp was.v called.v Edo.noun.	Tokyo.noun was.v formerly.adv called.v Edo.noun.

Q1: Lockjaw is another name for which disease? (tetanus)

Q2: Who wrote the opera Pagliacci? (leoncavallo)

Q3: What city was originally called edo? (tokyo)

Fig. 1. Example question and passage cases where part-of-speech of tokens in passages, named entities, and query term coverage are shown. Answer-containing passages demonstrate specific linguistic characteristics, e.g., more nominal terms, less pronouns, and sufficient query term coverage. Note: The questions and passages are taken from the QUASAR-T development set (see section 2.1 for more details).

Recent advances in fact-seeking QA and passage retrieval have been based on the utilization of distributed word representations as well as deep learning structures. The works in [9, 15] are based on the utilization of distributed word embeddings learned using word2vec [10] and GloVe [13] to represent the text of questions and passages with word-level embeddings and to find the most relevant passages. The work in [15] relies mainly on Convolutional Neural Networks and word embeddings as the discriminant analyzer to score and rank passages. The several rankers developed in [9] with LSTMs have been reported to outperform another deep learning-based passage retrieval system in [18] that developed a Reinforced Ranker-Reader QA system. The LSTMs were also employed in [2] in combination with word embeddings and character embeddings, and resulted in improvements over several baseline traditional and deep learning-based answer passage retrieval systems. The work in [11] made use of Bidirectional Encoder Representations from Transformers (BERT) [5]. The evaluation results of this technique on the TREC CAR and MS MARCO data sets show significant improvements over some of the state-of-the-art passage ranking techniques.

Beyond the distributed representation of characters and terms, there have been efforts to capture semantics of the larger portions of text, such as sentences [3, 8]. InferSent [3], for instance, developed sentence embeddings to encode the overall meaning of sentences. The InferSent encodings have been shown to generalize well to several natural language processing (transfer) tasks, such as multi-genre natural language inference and semantic textual similarity analysis. InferSent embeddings were used in [9] with feed-forward neural networks to train a passage ranking system which performed well on the QUASAR data set [6].

While previous works in the domain of answer passage retrieval and ranking have made significant progress in retrieving and ranking answer candidate passages at top ranks through the use of deep textual features (e.g., word and sentence embeddings), the possible contribution of more explicit utilization of linguistic features has not been studied.

Our approach to fill this gap is based on the utilization of both sentence-level semantics and explicit representation of several linguistic passage features. Focusing on passage ranking only and leaving aside answer extraction to further machine comprehension stages of QA that are not part of this work, we represent each sentence with its sentence embedding using InferSent [3]. The sentence embeddings are fed into a deep feed-forward neural network to predict whether or not a passage is likely to contain a candidate answer to a question. We then make use of the linguistic features of passages including token count, noun count, verb count, adverb count, pronoun count, query coverage, and named entity count to analyze the contribution of these features in passage ranking and to improve the final passage ranking effectiveness in terms of mean rank (MR) and answer recall of passages.

2 Methods

2.1 Data set

The QUASAR-T QA data set [6] was used in our experiments which includes training, development, and test subsets, each with short and long passages retrieved per question (100 short and 20 long passages). We focused on the short passages in the development and test subsets, each of which containing 3,000 questions. While there are other data sets for QA, e.g., SQUAD [14], the 1-to-many question-to-passage requirements are not met by such data sets to facilitate passage ranking experiments.

2.2 Deep neural network ranker

The baseline ranker in our analyses was a feed-forward deep neural network model. We constructed the input feature vector X_i to this ranker similar to the work in [9] and by concatenating question embedding (qe_i) and passage embedding (pe_i) that go through the network structure to find the answer-containing probability for passage i, as shown in the following equations.

Method	mr	r@1	r@2	r@3	r@4	r@5
BL-NN	9.58	0.25	0.39	0.47	0.54	**0.58**
R^3	n/a	**0.40**	n/a	0.51	n/a	0.54
InferSent	n/a	0.36	n/a	**0.52**	n/a	0.56

Table 1. MR and recall (r@top) analysis of our baseline model (BL-NN) and relevant methods on the test set. The results that are not available (not reported in referenced works) are shown with n/a.

$$X_i = [qe \oplus pe_i \oplus (qe - pe_i) \oplus (qe \odot pe_i)] \tag{1}$$

$$X_i^{flat} = \text{flatten}(X_i) \tag{2}$$

$$D^{(1)} = \text{ReLU}(W^{(1)} X_i^{flat}) \tag{3}$$

$$O_i = \text{softmax}(W^{(2)} D^{(1)}) \tag{4}$$

The embeddings for both questions and passages were constructed using InferSent where the output embedding per sentence has 4096 features. The input vector X, therefore, included 16,384 features. In cases where the text included more than one sentence, the embeddings of sentences were vector summed to create the representative sentence embedding of the entire text. To train this baseline model, the QUASAR-T development set of questions and short contexts were utilized. For each question, the contexts were first pre-processed and pseudo-labeled according to whether they contained the actual answer to the question. Then, a subset of 1 positive context and 5 negative contexts were extracted per question to train the baseline model in 500 epochs without any early stopping criteria or any regularization method. The pseudo-labeling of contexts resulted in a 2-feature output vector per context; hence, the output layer of the model is a dense layer including two neurons with Softmax activation. We modeled the passage ranking task as a classification problem similar to [11]; thus, the binary cross-entropy loss function was used, i.e., $L = -\sum_{i \in P_{pos}} log(p_i) - \sum_{j \in P_{neg}} log(p_j = 1 - p_i)$ where the P_{pos} and P_{neg} index sets represent the positive and negative pseudo-labeled contexts, p_i is the answer-containing probability (class=1), and p_j is the probability of class=0.

The trained model would then generate a probability per class, i.e., answer-containing versus answer-free. The answer-containing probability of each passage was used to rank passages. The loss and accuracy of the model in the training phase is shown in Figure 2. Table 1 shows the detailed results of this model when applied on the test set as compared with two existing, relevant (neural network-based) systems R^3 [18] and InferSent ranker [9] without the utilization of other lexical semantic features.

2.3 Explicit shallow linguistic features

Distributed sentence representations capture several surface, syntactic, and semantic characteristics of text [4]. However, in the context of QA, there are other

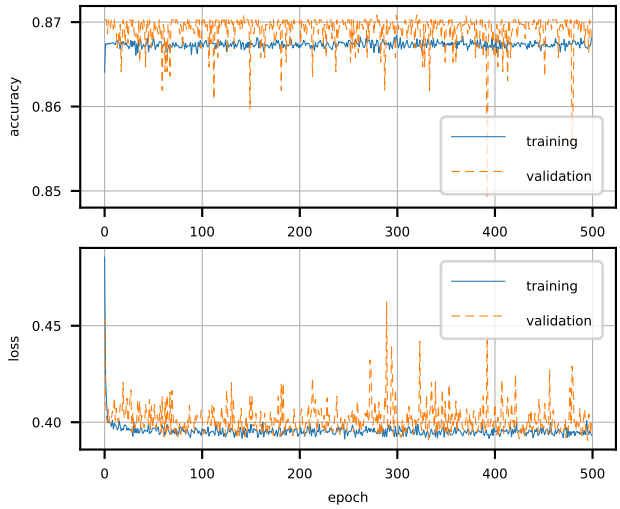

Fig. 2. Loss and accuracy analysis of the baseline neural network ranker on the QUASAR-T development set. Note: Higher validation accuracy values are due to the validation accuracy being calculated at the end of each epoch versus training accuracy being calculated batch-wise.

explicit linguistic features that can potentially enhance answer passage retrieval and ranking and yet they are under-represented in distributed embeddings. The explicit linguistic characteristics that we focused on included the following categories.

First, terms of specific part-of-speech can distinguish between an answer-containing passage and the one that is less likely to include an explicit answer to a question. These features include the number of nouns, verbs, adverbs, pronouns, as well as the general count of tokens within the text of a passage. It can be argued that for fact-seeking questions, it is less likely that the answer will be in the form of an adverb (or a verb) while it is more likely to be a noun or a nominal predicate. In addition, the larger number of tokens can be argued to have a positive impact on the chances of a passage containing the actual or candidate answer. Pronouns, on the other hand, can mask the actual answer within a passage and as such, the smaller number of pronouns may result in higher quality passages. As one example from the QUASAR-T data set, the question *"Which is considered the most powerful piece on the chess board?"* has contexts such as *"The queen is the most powerful piece in the board "* and *"She is the most powerful piece on the board"*. The correct answer is masked by the pronoun *"She"* in the second passage, which makes the passage less effective for QA purposes.

Second, in fact-seeking QA, the answer is most likely a text snippet that refers to the name of a location, organization, or a person. In some other cases, the

date, time, or a monetary reference is sought. While named entities have been used as a category of features for matching answer candidates and questions in [16], they have not been used for featurizing passages for their likelihood of answer-containing. We will show that a larger number of named entities are found in answer-containing passages.

Third, query term coverage within a passage was selected as another signal to correctly identify answer-containing passages. The argument is on the basis that correct actual answers to a given question may mostly be positioned in the close proximity of the same query terms that are mentioned in the text of the question. This is besides the fact that a larger proportion of query term coverage also contributes to the semantic relatedness of passages and questions. These two concepts (proximity of answers to query terms and coverage of terms) can be found in a more traditional passage retrieval and ranking system called MultiText [1].

We conducted an exploratory analysis of the above features in the QUASAR-T development set. The contexts were first pseudo-labeled; then, features were extracted for every context. There were 35,162 positive and 263,804 negative (answer-free) contexts. Separated by pseudo-labels (class=1 indicating answer-containing passages), Table 2 summarizes the descriptive statistics of the two cohorts of passages. The chart demonstrates that the medians and distributions of feature counts have meaningful differences between the two classes of passages in most cases. A two-tailed statistical t-test was then conducted on the distribution of each feature in the two passage classes and it was found that, except in the case of verb counts ($p = 0.31$), the means of all the other linguistic features were significantly different from each other at the 95% confidence level (with $p = 0.00$). The distributions show that answer-containing passages, on average, have larger token counts, noun counts, named entity counts, and query coverage while they also include smaller adverb and pronoun counts. These results were contradictory to one previous work in [17] which found that verbs can substantially contribute to the task of QA passage ranking. Our findings are, however, in agreement with the same work in terms of noun counts as [17] reported that nominal predicates can positively impact on answer passage ranking. As a result, we preserved all the explicit linguistic features in our experiments.

2.4 Fusion of linguistic features and deep semantics

2.4.1 Traditional machine learning fusion In the first attempt to enhance our baseline deep neural network ranker using explicit linguistic features, we used the answer-containing probability generated by the baseline ranker in combination with the explicit features extracted for each passage as the predictor set to re-classify the passages into positive versus negative classes. A number of traditional machine learning algorithms, including logistic regression, Gaussian naive Bayes, decision tree, random forest, linear support vector machines, and Sigmoid support vector machines, were utilized. To train each classifier, we applied our baseline neural network ranker on the QUASAR-T development data set to obtain the answer-containing probabilities (1 positive and 5 negative

Feature	class=0		class=1		p-value
	mean	stdv	mean	stdv	
noun count	8.44	6.27	10.94	6.36	0.00
verb count	2.33	1.86	2.32	1.77	0.31
token count	21.83	10.72	25.49	9.79	0.00
adverb count	0.55	0.87	0.53	0.84	0.00
named entity count	2.00	2.01	2.94	2.21	0.00
query coverage	1.86	1.18	2.36	1.25	0.00
pronoun count	0.56	0.97	0.42	0.81	0.00

Table 2. Descriptive analysis of linguistic features between answer-free (class=0) and answer-containing (class=1) contexts in the development data set.

Measure	LR	RF-bCV	SVM-sigmoid	SVM-linear	GNB	DT-bCV
AUC	0.59	0.81	0.53	0.48	0.56	0.58

Table 3. The AUC analysis of the second-level classification of passages using answer-containing probabilities of the BL-NN ranker and the explicit linguistic features.

contexts per question), where a Gaussian noise (mean=0.0 and standard deviation=0.1) was added to the probabilities for the baseline ranker was first trained using the same data set. Then, the linguistic features of the same development passages were extracted. These features and the answer-containing probabilities were then normalized using the $L2$ normalization technique and were fed into the machine learning techniques for training. Table 3 summarizes the AUC results of the different techniques, where the random forest and decision tree models went through a 5-fold cross-validation process to find the best maximum depth of the trees.

From the above machine learning techniques, the best random forest model found using cross validation (RF-bCV) had the best AUC; thus, it was selected for ranking of passages in the QUASAR-T test set. This model did not perform well as shown in Table 4.

2.4.2 Deep learning-based score and linguistic feature fusion A similar procedure to the traditional machine learning fusion approach (detailed in the previous section) was taken to train a deep feed-forward neural network model (with the same structure as in the baseline neural network ranker) in 50 epochs this time. The input to the second-level network (2^{nd}-NN) was low-dimensional and included the same answer-containing probabilities of the first baseline model (plus the Gaussian noise for training) as well as the linguistic features of passages; hence, the number of epochs was set to a much smaller number in this experiment (50 epochs). This model, when tested on the QUASAR-T test set, resulted in a better set of performance measures compared with the traditional random forest model in the previous experiment; however, the baseline neural network model was not improved upon as detailed in Table 4.

2.4.3 Deep learning-based augmentation In another experiment, the baseline deep neural network model was augmented with the explicit linguistic features extracted for passages. The augmentation of these features was done in the middle layer of the network by concatenating the outputs of the first dense layer (including 10 nodes) with the 7 linguistic features. The overall process, involving the augmented neural network ranker, is shown in Figure 3. The linguistic feature augmentation process is especially structured in the middle layer instead of the input layer with a large number of nodes (16,384) to more directly and strongly infuse the effect of the linguistic characteristics of passages into the neural model. The input vectors to this model include the same X_i in Equation 1 as well as LFs_i for linguistic features of passage i which go through the network structure to find the answer-containing probability of the passage as shown in Equations 5-7.

$$X_i^{flat} = \text{flatten}(X_i) \tag{5}$$

$$C^{(1)} = \text{ReLU}(W^{(1)} X_i^{flat}) \oplus LFs_i \tag{6}$$

$$O_i = \text{softmax}(W^{(2)} C^{(1)}) \tag{7}$$

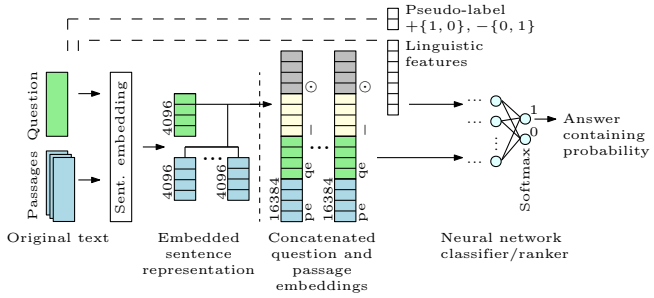

Fig. 3. The schematic view of the linguistically augmented passage ranking process for QA using a feed-forward deep neural network.

The augmented neural model was trained using the QUASAR-T development set with the same settings as in the baseline deep neural network model; a cross-entropy loss function, 500 training epochs, 1 positive passage, 5 negative passages, no early stopping criteria, and without drop-out or any other type of regularization. The loss and accuracy of the model in the training cycles were similar to those of the baseline ranker as shown in Figure 2.

More importantly, this model outperformed the baseline deep neural network ranker (BL-NN) with respect to all of the QA-based evaluation metrics in our experiments, i.e., MR and recall at different levels. The detailed results of this

Method	mr	r@1	r@2	r@3	r@4	r@5
BL-NN	9.58	0.25	0.39	0.47	0.54	0.58
RF-bCV	21.44	0.07	0.13	0.18	0.22	0.26
2^{nd}-NN	13.61	0.24	0.36	0.45	0.52	0.56
aug-NN	**8.79**†	**0.30**†	**0.44**†	**0.52**†	**0.59**†	**0.63**†

Table 4. MR and recall analysis of the rankers developed. aug-NN is the neural model augmented with linguistic features. The †s indicate statistically significant differences compared with BL-NN at the 95% confidence level.

Method	avg. loss	avg. accuracy
BL-NN	0.3624	0.8869
2^{nd}-NN	0.3580	0.8875
aug-NN	0.3610	0.8876

Table 5. Average loss and average accuracy analysis of the neural rankers developed on the test questions/passages.

model along with the other experimental models are summarized in Table 4. Although the improvements may seem marginal on the surface, the statistical test of significance on the large number of questions and passages in the benchmark data set proved otherwise. The statistical test was based on paired t-tests.

Our augmented neural model reached the performance of the best model in Table 1 (InferSent ranker) at r@3 and outperformed this model with respect to r@5 by a margin of 7%. It should be noted that the other comparison methods (R^3 and InferSent ranker) have relatively higher (base) performance values at r@$\{1, 3\}$ yet our proposed augmentation technique improves upon our weaker baseline model (BL-NN) to reach the performance of InferSent ranker at r@3 and significantly outperforms the two comparison methods at r@5. Also, while the proposed augmentation of shallow linguistic features was only applied on our BL-NN model and resulted in statistically significant improvements, a similar positive effect can be expected on the other comparison rankers too.

In terms of the classification performance of the several neural network models developed, the resulted of a detailed analysis of the average cross-entropy loss and average accuracy of the models are summarized in Table 5. These results are on the 3,000 QUASAR-T test questions and passages. As shown in Table 5, the average loss and accuracy of the models do not differ significantly (all within 1% variance); however, the QA-based metrics of final passage ranking have been shown to significantly improve using the augmented model.

3 Discussion

Passages that are more likely to contain specific answers to fact-seeking questions were shown to present with several linguistic features, mostly at the syntactic and lexical levels, that can further separate them from those that are less likely to recall any candidate answers. Even in presence of deep semantic

relatedness between questions and passages, surface and explicit features can eventually be assisting in distinguishing between positive answer-containing and negative passages and thus in better ranking of answer passages with a vision of improvements in overall QA effectiveness. The explicit lexical and syntactic characteristics of passages intrinsically increase chances of the text of a passage to contain a candidate or a correct answer to the question. The descriptive statistical analysis that we conducted on the two cohorts of pseudo-labeled passages (positive versus negative) along with statistical tests demonstrated significant differences between the distributions of the textual features in the two passage classes with the exception on verb counts. The latter finding regarding verbs is contradictory to the previous studies that showed verbs play a substantial role in answer passage retrieval [17].

In addition to the exploratory and descriptive statistical analysis of the surface linguistic features within passages, that suggest there are lexical differences between likely answer-bearing passages and those that are less likely to recall a candidate answer, the procedure taken to utilize these features was demonstrated to play an important role. The mere fusion of surface passage characteristics with answer-containing probability calculated through a more sophisticated deep semantic-oriented neural model was shown not to reach high levels of eventual answer passage ranking effectiveness measures. This failed experiment with both traditional machine learning and deep neural network models indicates that using the explicit linguistic features at a late stage of passage (re)classification and ranking is not effective.

To understand the relationship between the semantic relatedness measure of question-passage pairs and the explicit linguistic characteristics of passages, we used the answer-containing probabilities calculated for the contexts in the development set as a proxy for semantic relatedness and found the correlation between this measure and each of the linguistic features. We used the same set of 1 positive and 5 negative passages per question, the same data set that was used to train the second-level classifiers. As shown in Table 6, verb, adverb, and pronoun counts have the lowest correlations with the answer-containing probability of a passage, the latter two are negative. Noun and named entity counts have the largest (fair) correlations with the probability measure. None of the features were overly correlated with the probability of answer-containing, which removes the possibility of multicollinearity on answer-containing probability, and yet the method fails in better positioning answer passages.

The set of the same surface textual features combined internally within the structure of the deep neural network model (concatenated with the middle layer) fulfill the expectation of improvement over the effectiveness of the linguistic-feature-free baseline neural model. The statistically significant improvements over the performances of the baseline neural ranker support our hypothesis that the combination of the semantic relatedness of question-passage pairs (the output of the middle dense layer of the neural network model) and the surface passage features can improve answer passage ranking for fact-seeking QA.

Feature	#tokens	#nouns	#verbs	#adverbs	#pronouns	query coverage	#named entities
a.pr	0.48	0.57	0.01	-0.05	-0.14	0.22	0.55

Table 6. Correlation analysis between BL-NN answer-containing probabilities and the linguistic features of development passages. Note: a.pr=answer-containing probability.

The three neural network rankers we developed have very similar average cross-entropy loss and average accuracy values over the test questions and passages; however, in terms of the passage retrieval-based evaluation metrics (i.e., MR and recall), the ability of the rankers in positioning answer-containing passages at better ranks significantly differ from each other when linguistic features of passages are augmented within the network structure.

4 Conclusions

We analyzed the effect of several explicit, shallow linguistic features of textual passages that can enhance the overall effectiveness of answer passage ranking for fact-seeking QA. Several experiments were carried out to improve upon a baseline neural network ranker that makes use of deep semantics in sentence embeddings. The fusion of token count, noun count, verb count, adverb count, pronoun count, query coverage, and named entity count within passages with the answer-containing probabilities obtained through the application of the baseline neural model using traditional machine learning as well as a second-level neural network did not result in improved passage ranking effectiveness. However, when the same features were internally augmented with the middle layer of the baseline neural network ranker, the augmented model significantly outperformed the baseline ranker with respect to MR and recall at different levels. Our next steps will focus on more complex neural models and the effect of the infusion of a more comprehensive set of linguistic features, such as scenario-based and chunk-based textual relations as well as dependency trees/relationships.

References

1. Clarke, C.L.A., Cormack, G., Lynam, T., Li, C., McLearn, G.: Web reinforced question answering (MultiText experiments for TREC 2001) (2001)
2. Cohen, D., Croft, W.B.: A hybrid embedding approach to noisy answer passage retrieval. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR. Lecture Notes in Computer Science, vol. 10772, pp. 127–140. Springer (2018)
3. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Palmer, M., Hwa, R., Riedel, S. (eds.) EMNLP. pp. 670–680. Association for Computational Linguistics (2017)
4. Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M.: What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2126–2136.

Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1198
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) NAACL-HLT (1). pp. 4171–4186. Association for Computational Linguistics (2019)
6. Dhingra, B., Mazaitis, K., Cohen, W.W.: Quasar: Datasets for question answering by search and reading. CoRR **abs/1707.03904** (2017)
7. Goodwin, T.R., Harabagiu, S.M.: Medical question answering for clinical decision support. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. p. 297–306. CIKM '16, Association for Computing Machinery, New York, NY, USA (2016), https://doi.org/10.1145/2983323.2983819
8. Hill, F., Cho, K., Korhonen, A.: Learning distributed representations of sentences from unlabelled data. In: Knight, K., Nenkova, A., Rambow, O. (eds.) HLT-NAACL. pp. 1367–1377. The Association for Computational Linguistics (2016)
9. Htut, P.M., Bowman, S., Cho, K.: Training a ranking function for open-domain question answering. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 120–127. Association for Computational Linguistics, New Orleans, Louisiana, USA (Jun 2018). https://doi.org/10.18653/v1/N18-4017
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) ICLR (Workshop Poster) (2013)
11. Nogueira, R., Cho, K.: Passage re-ranking with BERT. CoRR **abs/1901.04085** (2019)
12. Oh, H.J., Myaeng, S.H., Jang, M.G.: Semantic passage segmentation based on sentence topics for question answering. Information Sciences **177**(18), 3696–3717 (2007). https://doi.org/https://doi.org/10.1016/j.ins.2007.02.038
13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
14. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad. CoRR **abs/1806.03822** (2018), http://arxiv.org/abs/1806.03822
15. Rosso-Mateus, A., González, F.A., y Gómez, M.M.: A two-step neural network approach to passage retrieval for open domain question answering. In: Mendoza, M., Velastin, S.A. (eds.) CIARP. Lecture Notes in Computer Science, vol. 10657, pp. 566–574. Springer (2017)
16. Suzuki, J., Sasaki, Y., Maeda, E.: SVM answer selection for open-domain question answering. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002)
17. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.A.: Using syntactic information for improving why-question answering. In: Scott, D., Uszkoreit, H. (eds.) COLING. pp. 953–960 (2008)
18. Wang, S., Yu, M., Guo, X., Wang, Z., Klinger, T., Zhang, W., Chang, S., Tesauro, G., Zhou, B., Jiang, J.: R3: Reinforced ranker-reader for open-domain question answering. In: AAAI (2018)
19. Wen, A., Elwazir, M.Y., Moon, S., Fan, J.: Adapting and evaluating a deep learning language model for clinical why-question answering. JAMIA Open **3**(1), 16–20 (2020). https://doi.org/10.1093/jamiaopen/ooz072, https://doi.org/10.1093/jamiaopen/ooz072