Fermi at SemEval-2019 Task 8: An elementary but effective approach to Question Discernment in Community QA Forums

Bakhtiyar Syed[1], **Vijayasaradhi Indurthi**[1,3], **Manish Shrivastava**[1],
Manish Gupta[1,2], **Vasudeva Varma**[1]
[1] IIIT Hyderabad, [2] Microsoft, [3] Teradata
[1]{syed.b, vijaya.saradhi}@research.iiit.ac.in
[1]{m.shrivastava, manish.gupta, vv}@iiit.ac.in
[2]gmanish@microsoft.com
[3]vijayasaradhi.indurthi@teradata.com

Abstract

Online Community Question Answering Forums (cQA) have gained massive popularity within recent years. The rise in users for such forums have led to the increase in the need for automated evaluation for question comprehension and fact evaluation of the answers provided by various participants in the forum. Our team, **Fermi**, participated in sub-task A of Task 8 at SemEval 2019 - which tackles the first problem in the pipeline of factual evaluation in cQA forums, i.e., deciding whether a posed question asks for a factual information, an opinion/advice or is just socializing. This information is highly useful in segregating factual questions from non-factual ones which highly helps in organizing the questions into useful categories and trims down the problem space for the next task in the pipeline for fact evaluation among the available answers. Our system uses the embeddings obtained from Universal Sentence Encoder combined with XGBoost for the classification subtask A. We also evaluate other combinations of embeddings and off-the-shelf machine learning algorithms to demonstrate the efficacy of the various representations and their combinations. Our results across the evaluation *test* set gave an accuracy of 84% and received the **first** position in the final standings judged by the organizers.

1 Introduction

The massive rise in popularity of Community Question Answering (cQA) forums like StackOverflow, Quora, Yahoo! Answers and Google Groups have led to an effective means of information dissemination for topic-centered communities to share and engage in knowledge consumption needs. After a considerable time, information becoming obsolete is a major problem which results in change of many of the facts that were previously true. Another problem is that most of the forums lack exhaustive moderation and control – which results in high-latency quality checks and eventually results in the sharing of non-factual information. Various factors are responsible for this – primarily being ignorance or misunderstanding and sometimes, maliciousness of the responder to the questions (Mihaylova et al., 2018).

In the pipeline of detection of whether the given responses to a question are indeed factual, the necessary first step is to discern what category the question asked in the cQA forum falls into. As an example, *"What is Domino's customer service number?"* is a factual question as it asks for a fact rather than an opinion or discourse. In contrast, consider the question *"Can someone recommend a good pediatrician in Mumbai?"* asks for an opinion rather than a particular factual information as opinions on the matter of a good pediatrician may be subjective and depend on various other factors the conclusion of which is not universally true.

We tackle the problem proposed by organizers (Mihaylova et al., 2019) in sub-task A as a multi-class classification problem, i.e., categorizing questions in cQA forums into one of the following three categories:

1. Factual: The question is asking for factual information, which can be answered by checking various information sources, and it is not ambiguous. *(e.g., "What is the currency used in Taiwan?")*

2. Opinion: The question asks for an opinion or an advice, not for a fact. *(e.g., "Can somebody recommend good restaurants around the SF Bay Area?")*

3. Socializing: Not a real question, but intended for socializing or for chatting. This can also mean expressing an opinion or sharing some information, without really asking anything of general interest. *(e.g., "What was your first bike?")*

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1160–1164
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

Our submission involves the use of pre-trained models for generating sentence embeddings from existing trained models and then employing the use of off-the-shelf machine learning algorithms for the multi-class prediction problem. The approach is described in Section 3 where we describe our methodology in detail.

2 Related Work

For classification tasks like question similarity across community QA forums, machine learning classification algorithms like Support Vector Machines (SVMs) have been used (Šaina et al., 2017; Nandi et al., 2017; Xie et al., 2017; Mihaylova et al., 2016; Wang and Poupart, 2016; Balchev et al., 2016). Recently, advances in deep neural network architectures have also led to the use of Convolutional Neural Networks (CNNs) (Šaina et al., 2017; Mohtarami et al., 2016) which perform reasonably well for selection of the correct answer amongst cQA formus. Algorithms and methods for answer selection also include works by (Zhang et al., 2017) which use a Long-Short Term Memory (LSTM) model for answer selection. Similarly, LSTMs for answer selection are also used by (Feng et al., 2017; Mohtarami et al., 2016). Other works in the space include use of Random Forests (Wang and Poupart, 2016); topic models to match the questions at both the term level and topic level (Zhang et al., 2014). There have also been works on translation based retrieval models (Jeon et al., 2005; Zhou et al., 2011); XgBoost (Feng et al., 2017) and Feedforward Neural Networks (NN) (Wang and Poupart, 2016).

All of the above related works on cQA used the features such as Bag of Words (BoW) (Franco-Salvador et al., 2016); Bag of vectors (BoV) (Mohtarami et al., 2016); Lexical features (for example, Cosine Similarity, Word Overlap, Noun Overlap, N-gram Overlap, Longest Common Substring/Subsequence, Keyword and Named Entity features etc.) (Franco-Salvador et al., 2016; Mohtarami et al., 2016; Nandi et al., 2017); Semantic features (for example, Distributed representations of text, Knowledge Graphs, Distributed word alignments, Word Cluster Similarity, etc.) (Franco-Salvador et al., 2016); Word Embedding Features (like Word2vec, GloVe etc.) (Wang and Poupart, 2016; Mohtarami et al., 2016; Nandi et al., 2017); and Metadata-based features (Mohtarami et al., 2016; Mihaylova et al., 2016; Xie et al., 2017).

In this work, we seek to evaluate pre-trained sentence embeddings and how they perform across comprehension of questions in the community QA tasks. We now describe the methodology and data in the following section.

3 Methodology and Data

The data supplied by organizers is used for the task at hand. Specifically, for sub-task A, the *subject* and *body* for each question is provided by the task organizers. The data consists of 1118 training instances along with 239 and 935 question instances in the development and testing sets respectively.

3.1 Word Embeddings

Word embeddings have been widely used in modern Natural Language Processing applications as they provide vector representation of words. They capture the semantic properties of words and the linguistic relationship between them. These word embeddings have improved the performance of many downstream tasks across many domains like text classification, machine comprehension etc. (Camacho-Collados and Pilehvar, 2018). Multiple ways of generating word embeddings exist, such as Neural Probabilistic Language Model (Bengio et al., 2003), Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and more recently ELMo (Peters et al., 2018).

These word embeddings rely on the distributional linguistic hypothesis. They differ in the way they capture the meaning of the words or the way they are trained. Each word embedding captures a different set of semantic attributes which may or may not be captured by other word embeddings. In general, it is difficult to predict the relative performance of these word embeddings on downstream tasks. The choice of which word embeddings should be used for a given downstream task depends on experimentation and evaluation.

3.2 Sentence Embeddings

While word embeddings can produce representations for words which can capture the linguistic properties and the semantics of the words, the idea of representing sentences as vectors is an important and open research problem (Conneau et al., 2017).

Finding a universal representation of a sentence which works with a variety of downstream tasks is the major goal of many sentence embedding techniques. A common approach of obtaining a sentence representation using word embeddings is by the simple and naïve way of using the simple arithmetic mean of all the embeddings of the words present in the sentence. Smooth inverse frequency, which uses weighted averages and modifies it using Singular Value Decomposition (SVD), has been a strong contender as a baseline over traditional averaging technique (Arora et al., 2016). Other sentence embedding techniques include p-means (Rücklé et al., 2018), InferSent (Conneau et al., 2017), SkipThought (Kiros et al., 2015), Universal Encoder (Cer et al., 2018).

We formulate sub-task A of Task 8 in SemEval 2019 as a text multi-classification task. In this paper, we evaluate various pre-trained sentence embeddings for identifying each of the categories of factual, socializing and opinion among the questions in community QA forums. We train multiple models using different machine learning algorithms to evaluate the efficacy of each of the pre-trained sentence embeddings for the sub-task. In the following, we discuss various popular sentence embedding methods in brief.

- InferSent (Conneau et al., 2017) is a set of embeddings proposed by Facebook. InferSent embeddings have been trained using the popular language inference corpus. Given two sentences the model is trained to infer whether they are a contradiction, a neutral pairing, or an entailment. The output is an embedding of 4096 dimensions.

- Concatenated Power Mean Word Embedding (Rücklé et al., 2018) generalizes the concept of average word embeddings to power mean word embeddings. The concatenation of different types of power mean word embeddings considerably closes the gap to state-of-the-art methods mono-lingually and substantially outperforms many complex techniques cross-lingually.

- Lexical Vectors (Salle and Villavicencio, 2018) is another word embedding similar to fastText with slightly modified objective. FastText (Bojanowski et al., 2016) is another word embedding model which incorporates character n-grams into the skipgram model of Word2Vec and considers the sub-word information.

- The Universal Sentence Encoder (Cer et al., 2018) encodes text into high dimensional vectors. The model is trained and optimized for greater-than-word length text, such as sentences, phrases or short paragraphs. It is trained on a variety of data sources and a variety of tasks with the aim of dynamically accommodating a wide variety of natural language understanding tasks. The input is variable length English text and the output is a 512 dimensional vector.

- Deep Contextualized Word Representations (ELMo) (Peters et al., 2018) use language models to get the embeddings for individual words. The entire sentence or paragraph is taken into consideration while calculating these embedding representations. ELMo uses a pre-trained bi-directional LSTM language model. For the input supplied, the ELMo architecture extracts the hidden state of each layer. A weighted sum is computed of the hidden states to obtain an embedding for each sentence.

Using each of the sentence embeddings we have mentioned above, we seek to evaluate how each of them performs when the vector representations of the body of questions in the cQA forums are supplied for classification with various off-the-shelf machine learning algorithms. For each of the evaluation tasks, we perform experiments using each of the sentence embeddings mentioned above and show our classification performance on the *dev* set given by the task organizers.

Model	**F-1**	**Acc**
Universal Encoder + XGB	**0.72**	**0.84**

Table 1: Results showing Macro-F1 score and accuracy for Sub-task A, using Universal Encoder Sentence embeddings and training the model with XGBoost.

4 Results

The official ranking metric is Accuracy. We have included the F-1 score here as well for comparison. Table 1 provides the results on the system runs for the evaluation phase as judged by the organizers on the CodaLab platform. Our system ranked *first* among the participants in the evaluation phase. We observe that Universal Sentence

Model	RF		SVM-RBF		XGBoost	
	Acc.	F-1	Acc.	F-1	Acc.	F-1
Universal Sentence Encoder	68.66	72.32	67.38	68.25	73.73	73.56
InferSent	53.91	50.89	61.56	63.45	60.82	59.32
Concat-p mean	56.22	49.01	65.64	69.54	60.36	60.01
Lexical Vectors	62.80	62.11	72.42	71.55	71.30	68.30

Table 2: *Dev* Set Accuracy and Macro-F-1 scores (in percentage) for **Sub-Task A** of Task 8

Encoder representations with the XGBoost classifier gives the best results on the test set.

As a way to elicit different performances for our experiments, we also provide our results from the system runs on the development set provided by the organizers. These results are shown in Table 2.

5 Conclusions and Future Work

We see from the results that our system is able to discern the type of questions asked in community QA forums with high performance metrics. This shows that using pre-trained embeddings with a simple machine learning classification algorithm often helps in greater understanding of the text at hand – in this case, the questions in community question-answering forums.

In future work, we also seek to evaluate different transfer learning approaches which utilize pre-trained language models (LMs) across different base language corpora and see how varying these base corpora for pre-training the language model results in the performance change while finetuning for question comprehension in cQA forums.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.

Daniel Balchev, Yasen Kiprov, Ivan Koychev, and Preslav Nakov. 2016. PMI-cool at SemEval-2016 Task 3: Experiments with PMI and Goodness Polarity Lexicons for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 844–850, San Diego, California. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Wenzheng Feng, Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2017. Beihang-MSRA at SemEval-2017 Task 3: A Ranking System with Neural Matching Features for Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 280–286, Vancouver, Canada. Association for Computational Linguistics.

Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. 2016. UH-PRHLT at SemEval-2016 Task 3: Combining Lexical and Semantic-based Features for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 814–821, San Diego, California. Association for Computational Linguistics.

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*, pages 84–90. ACM.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Tsvetomila Mihaylova, Pepa Gencheva, Martin Boyanov, Ivana Yovcheva, Todor Mihaylov, Momchil Hardalov, Yasen Kiprov, Daniel Balchev, Ivan Koychev, Preslav Nakov, Ivelina Nikolova, and Galia

Angelova. 2016. SUper Team at SemEval-2016 Task 3: Building a Feature-Rich System for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 836–843, San Diego, California. Association for Computational Linguistics.

Tsvetomila Mihaylova, Georgi Karadzhov, Atanasova Pepa, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez i Villodre, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James R. Glass. 2018. Fact checking in community forums. In *AAAI*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitra Mohtarami, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Tao Lei, Kfir Bar, Scott Cyphers, and Jim Glass. 2016. SLS at SemEval-2016 Task 3: Neural-based Approaches for Ranking in Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 828–835, San Diego, California. Association for Computational Linguistics.

Titas Nandi, Chris Biemann, Seid Muhie Yimam, Deepak Gupta, Sarah Kohail, Asif Ekbal, and Pushpak Bhattacharyya. 2017. IIT-UHH at SemEval-2017 Task 3: Exploring Multiple Features for Community Question Answering and Implicit Dialogue Identification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 90–97, Vancouver, Canada. Association for Computational Linguistics.

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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated p-mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Alexandre Salle and Aline Villavicencio. 2018. Incorporating subword information into matrix factorization word embeddings. *arXiv preprint arXiv:1805.03710*.

Filip Šaina, Toni Kukurin, Lukrecija Puljić, Mladen Karan, and Jan Šnajder. 2017. TakeLab-QA at SemEval-2017 Task 3: Classification Experiments for Answer Retrieval in Community QA. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 339–343, Vancouver, Canada. Association for Computational Linguistics.

Hujie Wang and Pascal Poupart. 2016. Overfitting at SemEval-2016 Task 3: Detecting Semantically Similar Questions in Community Question Answering Forums with Word Embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 861–865, San Diego, California. Association for Computational Linguistics.

Yufei Xie, Maoquan Wang, Jing Ma, Jian Jiang, and Zhao Lu. 2017. EICA Team at SemEval-2017 Task 3: Semantic and Metadata-based Features for Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 292–298, Vancouver, Canada. Association for Computational Linguistics.

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 Conference on Information and Knowledge Management*, pages 371–380. ACM.

Sheng Zhang, Jiajun Cheng, Hui Wang, Xin Zhang, Pei Li, and Zhaoyun Ding. 2017. FuRongWang at SemEval-2017 Task 3: Deep Neural Networks for Selecting Relevant Answers in Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 320–325, Vancouver, Canada. Association for Computational Linguistics.

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: Human Language Technologies-Volume 1*, pages 653–662. Association for Computational Linguistics.