A Deep Learning Methodology for Semantic Utterance Classification in Virtual Human Dialogue Systems

Debajyoti Datta, Valentina Brashers, John Owen, Casey White, and Laura E. Barnes

University of Virginia, Charlottesville, VA
{dd3ar,vlb2z,jao2b,cw4xz,lb3dp}@virginia.edu

Abstract. This paper describes the development of a deep learning methodology for semantic utterance classification (SUC) for use in domain-specific dialogue systems. Semantic classifiers need to account for a variety of instances where the utterance for the semantic domain class varies. In order to capture the candidate relationships between the semantic class and the word sequence in an utterance, we have proposed a shallow convolutional neural network (CNN) along with a recurrent neural network (RNN) that uses domain-specific word embeddings which have been initialized using Word2Vec for determining semantic similarity of words. Experimental results demonstrate the effectiveness of shallow neural networks for SUC.

Keywords: dialogue systems, interprofessional medical education, intelligent virtual agents, healthcare

1 Introduction and Related Research

Recent progress in deep learning approaches have transformed fields such as natural language processing. In particular, these new advances create new opportunities in the field of intelligent virtual agents (IVA). One of the key components of IVA systems is the dialogue system. Dialogue systems aim to automatically identify the intent of the user as expressed in natural language, and then perform the corresponding task specific to the domain.

The majority of the work in dialogue systems relies on semantic utterance classification for the evaluation of natural language queries into a particular category and then determining the updates to the dialogue states [13]. Typically, these systems use supervised classification methods like boosting [10], support vector machine approaches [12] or maximum entropy models [14]. In this work, we propose techniques for automated feature engineering using deep learning and task specific word embeddings. We improve upon existing approaches in which feature engineering is often task specific and cannot be generalized to different domains, thus limiting their reuse. Our work aims to create a reusable framework for domain-specific intelligent virtual agents [7]. We present the utility of our

proposed approach in the context of an IVA-delivered medical interprofessional education scenario.

2 Approach

Target Scenario. We focused the proposed approach on an IVA-delivered medical interprofessional education training scenario aimed at improving communication among members of healthcare teams. In the scenario, a nursing student must perform an assessment of a virtual patient with chronic obstructive pulmonary disease (COPD) exhibiting shortness of breath. The student must engage in teamwork communication with a virtual medical provider and using a validated checklist of fifteen behaviors called the Collaborative Behaviors Observational Assessment Tools (CBOATs) [1]. Figure 1 depicts a subset of the SUC categories along with sample user statements and virtual agent responses.

In domain-specific dialogue systems, like the one in this scenario, intent determination is the key element. Previously used intent determination approaches require heavy feature engineering [14, 10] over multiple iterations which slows down the building of an end-to-end SUC system and also limits the reuse of existing systems. We have proposed a deep learning approach that does not rely on task-specific feature engineering and can be rapidly trained and deployed on various scenarios. The deep learning approach has three key components: 1. Word Embeddings, 2. Convolutional Neural Network (CNN) for local and global semantics, and 3. Recurrent Neural Network (RNN) to capture word dependencies.

Word Embeddings. The fundamental idea behind Word2Vec is the distributional hypothesis, i.e words are characterized by the company that they keep. CBOW and Skipgram [8] are the main approaches for the learning the word embeddings. Since words are the atomic unit of each sentences, each sentence has different representations based on the context in which it is used. In the case of the COPD patient, the two sentences, "How are you doing today?" and "Hi John, how are you feeling now?" mean the same thing, even though in another context they may have a different meanings. Thus, training word embeddings [8, 9] improves classification accuracy over traditional indices based approaches.

Convolutional Neural Networks and Recurrent Neural Networks. CNNs popularized by Lecun [6] for image classification have since been used for a wide variety of NLP tasks like semantic parsing [15], search query retrieval [11], sentence modeling [5], and other traditional NLP tasks [3]. Traditionally in natural language processing when the task is to predict based on ordered set of items like words in a sentence or sentences in a document, some items convey more information than others. Order and the position of the words are also important in determining the meaning of the sentences. For example the two sentences, "It was not good, it was actually quite bad" and "It was not bad, it was actually quite good", have the same words, but different ordering and completely opposite intents. In cases like this, bag-of-words or n-grams will not work very

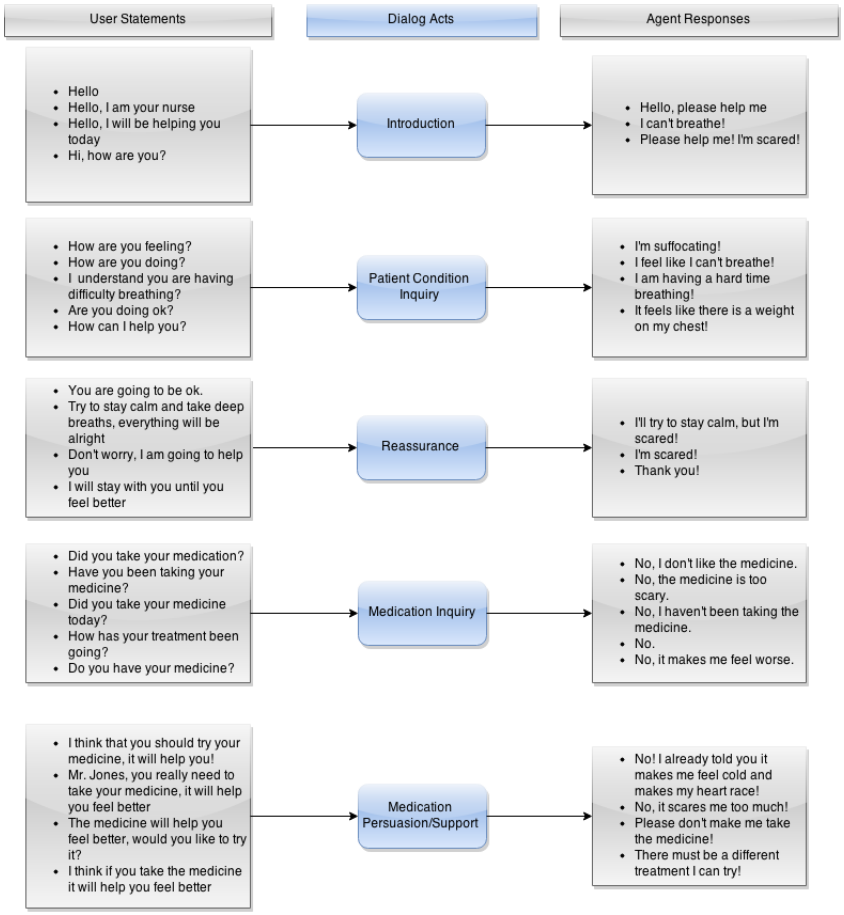

Fig. 1. IPE scenario and sample dialogue acts.

effectively or will result in huge and sparse embedding matrices. In this case, CNN architectures work particularly well and are a robust and elegant solution to the problem. CNNs utilize layers with convolving filters that are applied to local features. In the proposed implementation of CNN, there is one layer of convolution applied on top of the word embedding vectors. The convolution operation involves multiple filters which is applied to a window of words that produces a new feature.

Recurrent neural networks (RNN) are good at modeling temporal word dependencies. For example in the sentence, "I was born in France and lived there for the last 18 years and therefore I speak fluent French", in order to predict, the last word, "French" from the previous words in the sentence, one would have to observe dependencies that occurred much earlier in the sentence. We use two RNN architectures, Long-Short Term Memory (LSTM) [4] and Gated Recurrent

Unit (GRU) [2], to model these long-term dependencies. The proposed approach utilizing both, CNNs and RNNS, is capable of more fine-grained analysis and distinction.

Results. To evaluate the proposed approach, we transcribed 54 videos of nursing students interacting with standardized patients in the target COPD scenario and coded sentences according to the CBOAT categories shown in Figure 1. Each of these videos had roughly 20 interactions with a total of 2300 sentences. The convolution operation involves multiple convolving filters which is applied to a window of words which produces a new feature. Each filter is applied to each possible window of words in the sentence to produce a feature map and then max pooling over time operation is applied over the feature map to take the maximum value as the feature. Thus, because of padding and then taking the max pool operation, this method can easily deal with sentences of any length. The filter lengths can be varied along with dropout probabilities for regularization, and the training is done with the ADADELTA update rule [16]. We evaluated the proposed approach using 10-fold cross validation using random, static, and non-static word embeddings with various network architectures. Results demonstrate the effectiveness of the proposed approach for semantic utterance classification in domain-specific dialogue systems achieving an overall accuracy of 96.3% with the CNN, simple RNN, and non-static word embeddings. Table 1 shows the results for other architectures.

Table 1. Performance across different network architectures on IPE data set

	CNN	CNN + SimpleRNN	CNN + GRU	CNN + LSTM
Word2Vec non-static	90.1%	96.3%	92.1%	93.9%

3 Conclusion and Future Work

We proposed a deep learning methodology specifically targeted at intent determination tasks. In the proposed method, the random or pre-trained word embeddings are fed into a recurrent neural network (LSTM, GRU or a Simple RNN) to capture dependencies among words in the sentences. The output from the RNN is then fed into multi-channel convolutional layers to capture local semantics. The max over time pooling layers capture global semantic features followed by a fully connected layer with dropout to summarize the features. Preliminary experiments demonstrate that the approach outperformed traditional feature engineered approaches for intent determination tasks. As future work, we plan to benchmark our approach on other data sets and test our system with real users.

Acknowledgments. This research was supported in part by an Ivy Foundation Biomedical Innovation Grant.

References

1. Brashers, V., Erickson, J.M., Blackhall, L., Owen, J.A., Thomas, S.M., Conaway, M.R.: Measuring the impact of clinically relevant interprofessional education on undergraduate medical and nursing student competencies: A longitudinal mixed methods approach. Journal of Interprofessional Care 30(4), 448–457 (2016), http://dx.doi.org/10.3109/13561820.2016.1162139, pMID: 27269441
2. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
3. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. The Journal of Machine Learning Research 12, 2493–2537 (2011)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
5. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188 (2014)
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
7. Mancini, M., Ach, L., Bantegnie, E., Baur, T., Berthouze, N., Datta, D., Ding, Y., Dupont, S., Griffin, H.J., Lingenfelser, F., Niewiadomski, R., Pelachaud, C., Pietquin, O., Piot, B., Urbain, J., Volpe, G., Wagner, J.: Laugh when you're winning pp. 50–79 (2014), http://dx.doi.org/10.1007/978-3-642-55143-7_3
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
9. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation.
10. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. Machine learning 39(2), 135–168 (2000)
11. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural networks for web search. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 373–374. WWW '14 Companion, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2567948.2577348
12. Silva, J., Coheur, L., Mendes, A.C., Wichert, A.: From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review 35(2), 137–154 (2011)
13. Tur, G., De Mori, R.: Spoken language understanding: systems for extracting semantic information from speech. John Wiley & Sons (2011)
14. Tur, G., Deng, L., Hakkani-Tür, D., He, X.: Towards deeper understanding: Deep convex networks for semantic utterance classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5045–5048. IEEE (2012)
15. Yih, W.t., He, X., Meek, C.: Semantic parsing for single-relation question answering. In: Proceedings of the ACL. pp. 643–648 (2014)
16. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)