Vector Representations of Idioms in Chatbots

Tosin P. Adewumi, Foteini Liwicki, and Marcus Liwicki

Luleå University of Technology, Luleå, Sweden — firstname.lastname@ltu.se

Abstract. Open-domain chatbots have advanced but still have many gaps. My PhD aims to solve a few of those gaps by creating vector representations of idioms (figures of speech) that will be beneficial to chatbots and natural language processing (NLP), generally. In the process, new, optimal fastText embeddings in Swedish and English have been created and the first Swedish analogy test set, larger than the Google original, for intrinsic evaluation of Swedish embeddings has also been produced. Major milestones have been attained and others are soon to follow. The deliverables of this project will give NLP researchers the opportunity to measure the quality of Swedish embeddings easily and advance state-of-the-art (SotA) in NLP.

Keywords: NLP · Embeddings · Swedish analogy set · Chatbots · Idioms.

1 Introduction

Artificial Intelligence (AI) in conversational agents (chatbots) began with Joseph Weizenbaum's ELIZA in the 1960s [6]. Despite advancements recorded over the decades, there exist gaps, one of which is the inability of such agents to understand idioms in natural languages. My project addresses this problem under the title "Vector Representations of Idioms in Data-Driven Chatbots for Robust Assistance". In order to solve the challenge, two research questions are considered:

- To what extent can idioms embeddings enhance natural language processing in English, Swedish and Yoruba languages?
- How can idioms embeddings enhance open-domain, data-driven chatbots for robust assistance?

Embeddings (vector representations) are useful in neural network architectures in NLP for downstream tasks, which agents utilize. fastText by Grave et al [4] is currently adopted to create word embeddings in both Swedish and English. It has the advantages of speed and competitive performance to SotA. Although many different language embeddings have been created by researchers, many analogy test sets for evaluation do not exist, including one for the Swedish language [2, 4], because they can be expensive to create. Therefore, our immediate objectives and contributions are to create and evaluate optimal Swedish and English embeddings for NLP purposes, as well as create the first Swedish analogy

test set, similar to the Google set, for intrinsic evaluation[1]. There is also ongoing effort to create idioms datasets.

2 Methodology

SotA architectures like the Transformer will be trained on applicable datasets and the embedding layer provided by pre-trained embeddings. The training datasets used for the English and Swedish embeddings are 2019 Wikipedia dumps of 27G and 4G, respectively, after pre-processing [7, 8]. Both embeddings were generated using the original C++ fastText implementation [4], based on optimal hyper-parameters determined from previous work [1]. We ran the experiments on a shared DGX cluster with 80 GPUs on Ubuntu 18 operating system.

2.1 Datasets

Besides the Wikipedia datasets for unsurpervised training for the embeddings and the analogy test sets identified earlier, others required for the project are the labelled idioms datasets in English, Swedish and Yoruba languages and an analogy test set for Yoruba. The planned datasets will be the largest in the various languages and available for researchers.

3 Results and Deliverables

It is planned that we publish our research in peer-review journals and build a data-driven, open-domain chatbot implementing our research. Our results from the embeddings are competitive to the current state-of-the-art. This is based on intrinsic evaluations of the English models, evaluated on the Google analogy and WordSimilarity-353 (and corresponding Spearman correlation) test sets [3, 5] and evaluations of the Swedish models using the newly created analogy test set. The second level of evaluation of these pre-trained embeddings will be in downstream NLP applications related to the chatbot.

4 Conclusion

In conclusion, the PhD project involves many steps, of which some have been accomplished and some ongoing, in accordance with our Gantt chart. New, optimal fastText embeddings, based on previous research, in Swedish and English have been created. Importantly, the first Swedish analogy test set (of over 20,800 samples), larger than the Google original for intrinsic evaluation of Swedish embeddings has also been produced. There's ongoing and future work in producing idioms datasets in the relevant languages to be trained on SotA neural network architecture for deployment in open-domain chatbots as the last leg of the project.

[1] The test set is available, under the same licence, upon publication of the work

References

1. Adewumi, T.P., Liwicki, F., Liwicki, M.: Word2vec: Optimal hyper-parameters and their impact on nlp downstream tasks. arXiv preprint arXiv:2003.11645 (2020)
2. Fallgren, P., Segeblad, J., Kuhlmann, M.: Towards a standard dataset of swedish word vectors. In: Sixth Swedish Language Technology Conference (SLTC), Umeå 17-18 nov 2016 (2016)
3. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. ACM Transactions on information systems **20**(1), 116–131 (2002)
4. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893 (2018)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
6. Weizenbaum, J.: Eliza—a computer program for the study of natural language communication between man and machine. Communications of the ACM **9**(1), 36–45 (1966)
7. Wikipedia: English wikipedia multistream articles (2019), https://dumps.wikimedia.org/backup-index.html
8. Wikipedia: Swedish wikipedia multistream articles (2019), https://dumps.wikimedia.org/backup-index.html