Project Proposal

Joshua Rivera, Karan Sagar, Evan Silverman
LING-UA 52 Spring '21

Motivation

Despite existing research (Kiros '15, Conneau '17, Cer '18) there remains room to explore the semantic regularity of sentence embeddings. Evaluating regularities can be challenging due to the lack of a clear inverse function for pre-trained language models. Some approaches include probing (Conneau '18) and natural language generation (Kerscher '20, Wang '20).

Until recently, VAEs (Kingma and Welling '13, Bowman '16), which have a built-in correspondence from the embedding space to the sentence space, have not shown comparable performance on GLUE tasks. However, with OPTIMUS (Li '20) does and, as a result, we believe exploration of semantic regularity via syntactic analogy -- a:b::c:d for sentences a, b, c, d -- using OPTIMUS could prove fruitful.

Plan of Work

We plan to use the pre-trained OPTIMUS model across a variety of analogy types. To be explicit, for an analogy $a: b:: c: d$ and corresponding embeddings S_a, S_b, S_c, S_d, we are solving for S_d given S_a, S_b, S_c based on the equation $S_d \approx S_b - S_a + S_c$

This includes lexical, syntactic, and relationship analogies. For examples of the first two, please see Zhu '20. Relationship analogies refer to identical relationships from an NLI dataset: "The turtle is tracking the fish":"The turtle is following the fish"::"A person is dicing an onion":"A person is cutting an onion to pieces". Notice that entailment is the shared NLI relationship; the same will apply to negation. NLI data labeled as "neutral" will not be included

To measure performance on our evaluation set, we plan on using:

- Levenshtein distance
- BLEU score
- Proportion of evaluation data that is identically generated.

Tools / Requirements

For this project, we will use the pre-trained OPTIMUS model (https://github.com/ChunyuanLI/Optimus).

For data, we plan to use the Zhu '20 dataset methodology, which replaced works according to a specific template. If we can, we will also supplement with the data from that paper directly. For relationship analogies, we will use existing NLI datasets, including Multi-NLI (Williams 18) and SNLI (Bowman 15).

Data Collection

- Zhu '20 dataset
- Multi-NLI: https://cims.nyu.edu/~sbowman/multinli/

Collaboration statement

All team members participated in developing the core ideas. Joshua and Evan connected with Alex, Karan helped do the initial research and clarified concepts. All members participated in the writing of this document.

Citations

- Xunjie Zhu, Gerard de Melo. 2020. Sentence Analogies: Linguistic Regularities in Sentence Embeddings. *International Committee on Computational Linguistics* 3389-3400

- Mikolov, et. al. 2013 (a). Efficient Estimation of Word Representations in Vector Space (word2vec). arXiv manuscript 1301.3781
- Jeffrey Pennington, Richard Socher, Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. *Association for Computational Linguistics* 1532-1543
- Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig. 2013 (b). Linguistic Regularities in Continuous Space Word Representations. *Association of Computational Linguistics* 746-751
- Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, Jianfeng Gao. 2020. Optimus: Organizing Sentences via Pre-trained Modeling of a Latent Space. arXiv manuscript 2004.04092
- Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. arXiv manuscript 1606.07736
- Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard Zemel, Antonio Torralba, Raquel Urtasun, Sanja Fidler. 2015. Skip-Thought Vectors. arXiv manuscript 1506.06726
- Hsiao-Yu Chiang, Jose Camacho-Collados, Zachary Pardos. 2020. Understanding the Source of Semantic Regularities in Word Embeddings. *Association for Computational Linguistics* 119-131
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. arXiv manuscript 1705.02364
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. arXiv manuscript 1805.01070
- Adina Williams, Nikita Nangia, Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Association for Computational Linguistics* 1112-1122
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. arXiv manuscript 1508.05326
- Diederik P Kingma, Max Welling. 2013. Auto-Encoding Variational Bayes. arXiv manuscript 1312.6114
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, Samy Benigio. 2016. Generating Sentences from a Continuous Space. arXiv manuscript 1511.06349
- Daniel Cer, et. al. 2018. Universal Sentence Encoder. arXiv manuscript 1803.11175
- Martin Kerscher, Steffen Eger. 2020. Vec2Sent: Probing Sentence Embeddings with Natural Language Generation. arXiv manuscript 2011.00592
- Liyan Wang, Yves Lepage. 2020. Vector-to-Sequence Models for Sentence Analogies. *International Conference on Advanced Computer Science and Information Systems*