Topic: Sentence Similarity Detection

Paper #1: Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. *Deep Contextualized Word Representations*. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Summary: Previous deep contextualized word representation methods did not contain enough high-level (contextual) understanding, instead mainly focusing on syntax and word structure. With introduction of high-level context evaluation, can increase effectiveness on data sets that include challenging material, such as sentiment analysis and question/answer rapport. If able to better compute context, high-quality interpretations become more accessible to current models simply with the addition of a high-level word representation model. Previous work that impacted this field includes Liu et al. (2017), which set the state of the art on a majority of the datasets used within the paper. The paper seeks to improve upon the model used in the previous, and by adding contextual understanding increase the scores on NLP problems. Datasets used include the Stanford QuestionAnswering Database (SQuAD), CoNLL 2003/2012 Corpuses and the Stanford Sentiment Tree Bank(SST-5) to improve and evaluate the performance of NLP Tasks such as semantic role labelling and textual entailment. By developing and implementing a high-level contextual structure with pre-existing low-level syntax-based structures, a more expansive network was created that better dealt with problems involving context and indirect links between questions and answers. The ELMo model put forward used functions that represented entire sentences, instead of individual words. Their use of bidirectional language models (biLMs) underneath this sentence-vector structure returns a model that can both understand an input piece of data at the single word level and the overall sentence contextual level (high level). The use of the ELMo model proved ineffective for improved performance in textual entailment. This form of deductive reasoning did not seem to be aided by the addition of high-level contextual calculations. Improving the hypothesis-prediction portion of evaluation would require the improvement of higher level structures.

Paper #2: Liu, X., Shen, Y., Duh, K., & Gao, J. (2018). *Stochastic Answer Networks for machine reading comprehension*. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/p18-1157

Summary: This paper seeks to address the shortcomings of previous Stochastic Answer Networks (SAN) with respect to the inference of an answer based on subtle and complex clues. The primary implementation in this design is the use of a “dropout” layer on the final step of the answer module for training the DNN. This method drops certain predictions based on external weightings provided y the dataset in order to better smooth out the “jump” from reasoning to answer prediction. By increasing the accuracy of these jumps, SAN models can become more efficient at dealing with fringe cases of logical reasoning with respect to classification of both question and answer, and inference data. By building on SANs that were developed from 2012-2016, the paper seeks to improve the both the overall score on well-established data sets, as well as increasing the efficiency of the model in evaluating specific “vague” data. The use of SQuAD is similar to the previous paper

mentioned above, with this iteration attempting to better answer questions that require a logic "jump". MS MARCO is a unique dataset that is composed of real-world Bing Q + A's that better train the DNN in dealing with logic jumps. By employing two major layers (the Contextual Encoding Layer and the Memory Generation Layer), the model is able to effectively "break up" the received data into multiple different possible interpretations, and with the use of the "dropout layer" in the training process, certain approaches are weighted out of the final behaviour. The SAN model outperformed all previous stochastic models in the field, by a significant margin. The dropout layer assisted in improving this score, as a majority of false positives that plagued the previous models were reduced as a result of this logic elimination. One shortcoming of the model, however, is that it is simplistic in comparison with non-stochastic models. This means that by possibly implementing elements from other models, the SAN network can begin to compete with non-stochastic models (which for the most part out-scored the SAN model on straightforward portions of the dataset).

Paper #3: Hochreiter, S., Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9, 1735–1780.

(Dataset application for this foundational model can be found in:
Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.2). PhysioNet.
https://doi.org/10.13026/zx4k-te85.)

Summary: Considered a foundational paper in the development of sentence interpretation DNNs, this paper from 1997 seeks to lay out an O(1) complexity recurrent neural network that minimizes complexity associated with storing data. By creating such an efficient storage mechanism, the development of task-oriented neural networks can be done. The importance of this paper revolves around giving neural network architects the ability to manufacture their networks with quick gradient based recall methods that allow the sufficient training needed to employ a neural network with high accuracy. The use of LSTM is widely seen, including fields such as healthcare and social media categorization. The writers of the paper did not use any real-world data sets to evaluate the model, as a result of the period which the paper was written. Yet, the method has been applied after the fact on datasets such as PTB-XL an electrocardiography data set that is regularly employed on medical diagnoses-oriented neural networks. This dataset evaluation was successful (Wagner Et al., 2022), demonstrating the effectiveness of such a foundational model. LSTM removes the issue of a vanishing gradient for backpropagation networks by refusing to change gradients when not accessed. By doing this, the system avoids the tendency of gradients to go to 0 with regards to their weighting criteria. With this approach, the method outlined allowed for a whole new set of problems to be immediately solved as a result of its extremely efficient use of complexity.

Comparison Table:

Paper	Problem Addressed	Data Used	Models Implemented	Results	Shortcomings
Peters et al., (2018)	Added additional high level functionality to understand contextual variation, while still maintaining syntactic understanding.	SQuAD SNLI OntoNotes CoNLL 2003 CoNLL 2012 SST-5	Embeddings from Language Models (ELMo)	Placed highest scores at time of publication for all 6 datasets used, with minimum error reduction of 0.7% and maximum of 4.7% (absolute).	Did not significantly improve in the area of textual entailment, due to a limited ability to "jump" logic gaps. This was later addressed by Paper #2 in this literature review.
Liu et al., (2018)	Inability for previous SANs to infer answers based on the extraction of multi-step reasoning indicators.	SQuAD MS MARCO	Stochastic Answer Network (SAN) with dropout layer in training.	Outperformed all previous stochastic models, but had similar results compared to non-stochastic models of the period.	Failed to outperform non-stochastic models with respect to non-complex portions of the evaluated dataset. This is due to the simplicity of the surrounding model, yet by including elements of non-SANs, the efficiency of this model can be increased in theory.
Hochreiter et al., (1997)	Developing an O(1) complexity information storage system using a gradient approach for a back-propagation network.	None at time, later applied to datasets such as PTB-XL	Long Short-Term Memory (Gradient Based)	Became a foundational structure for later development in neural network development, being one of the most cited papers on gradient-based backpropagation.	Was not originally applied to datasets, as relevant ones were not available yet. However, it was recently applied to medical datasets with a high degree of success.

Sample Data

Dataset Name	Size	Type of Data and Features (columns)	# of Training/Testing samples
Stanford Question Answering Database 1.1 (SQuAD)	100K	Question and Answer (Wikipedia Responses to Common Questions. 1.1 Iteration contains: -Title (Topic of Question) -Question(s) -Answer (From Wikipedia Paragraph) Example Data: {paragraph= "[Wikipedia] In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity.**", question = "What causes rain to fall?", answer = "**gravity**".}	100K+ at the time of publication, with crowdsourced Q/As coming in throughout development. Utilized in the *Question Answering* portion of Peters Et al. (2018). **85.8%** Utilized in Liu Et al. (2018) **84.0%**
Stanford Natural Language Inference (SNLI) corpus	570K	Sample Data: {text = "A soccer game with multiple males playing.", judgements = "entailment \n E E E E E", hypothesis = "Some men are playing a sport."}	570,000 human-labelled sentence pairs Utilized in the *Textual Entailment* portion of Peters Et al. (2018). **88.7%**
OntoNotes 5.0 corpus	1.45M English Sentences	Spans several genres in 3 languages (English, Chinese, Arabic). Contains structural and semantic information for the utilization in DNN Semantic Role Labelling Sample Data: {plain = "I ground the rye on number 6 click-LRB-out of 8-RRB-in my Champion Juicer grinder." Leaves = ["I","ground","the","rye","on","number"]}	1,445,000 English, 300,000 Arabic, 900,000 Chinese sentences. Utilized in the *Semantic Role Labelling* portion of Peters Et al. (2018). **84.6%**
CoNLL-2003 NER Task	300K+ Tokens	Reuters news stories between 1996-1997. 4 Columns space separated: Word, Part-of-Speech(PoS), syntactic chunk, named entity tag. Sample Data:	Collection of words, sentences and tags from Reuters news stories utilized in order to test on named entity extraction.

		“ U.N. NNP I-NP I-ORG official NN I-NP O Ekeus NNP I-NP I-PER heads VBZ I-VP O for IN I-PP O Baghdad NNP I-NP I-LOC . . O O “	Utilized in the *Coreference Resolution* portion of Peters Et al. (2018). **70.4%**
CoNLL-2012 Shared Task	1M+ Sentences 200K Bible Translations	Collection of both news stories (categorized) and New Testament translations. Sample Data: “Argentina said it will ask creditor banks to *halve its foreign debt of $64 billion — the third-highest in the developing world . Argentina aspires to reach *a reduction of 50% in the value of its external debt.”	Utilized in the *Named Entity Extraction* portion of Peters Et al. (2018). **92.22%**
Stanford Sentiment Tree Bank (SST-5)	11,855 Movie Review Sentences	Composed of movie review sentences drawn from websites, and manually labelled with the following categories: -negative, positive, somewhat negative, somewhat positive Sample data: “This movie doesn't care about cleverness, wit or any other kind of intelligent humor. Those who find ugly meanings in beautiful things are corrupt without being charming. There are slow and repetitive parts, but it has just enough spice to keep it interesting.”	Utilized in the *Sentiment Analysis* portion of Peters Et al. (2018). **54.7%**
Microsoft Machine Reading Comprehension Dataset (MS MARCO)	100K	Bing queries from the real world. Sample data: [“Approximately 16,000 per year”, id = 2713, is_selected = 1, passage_text = “How many puffins are born in south Africa?”]	Utilized in Liu Et al. (2018) **46.14%**