Do Language Models Understand Sentences? An Analysis in Paraphrase Identification

Yunyi Xu[1], Xiangyu Liu[2], Xinyuan Mo[3]

I. Abstract

This study investigated whether the 3 language models of different complexities (Bag-of-n-grams, MPNET, and BERT) actually capture the semantics or the statistical features of language forms, by a set of designated paraphrase identification experiments. 2 datasets consisted of paraphrases or non-paraphrases pairs, namely Quora Question Pairs (Quora) and Paraphrase Adversaries from Word Scrambling (PAWS)[1], are used to carry out the experiments. The corpus of Quora is statistically much easier than that of PAWS. Therefore, we made the hypothesis that models that can capture semantics would perform roughly equal with both datasets. Our experiments have shown that all 3 models performed significantly poorer on PAWS than Quora, suggesting that they are not powerful enough to capture the precise semantic textual features and understand the meaning of the sentences.

II. Introduction

Paraphrase identification (PI) is concerned with the ability of identifying alternative linguistic expressions of the same meaning at different textual levels (document level, paragraph level, sentence level, word level, or combination between them)[2], which is of the most popular NLP tasks today. With the development of the large neural language model such as BERT[3] and MPNet[4], there has been tremendous progress on a wide range of NLP tasks, especially those are meaning-sensitive (just like PI).

However, some concerns are related to the deeper and deeper neural networks. Bender and Koller (2020)[5] doubted the claims that language models such as BERT can "understand" natural language or learn its "meaning".

In this work, we investigate the performances of language models of various complexities, and the their abilities to understand different language models on the PI tasks of two different datasets. If a model can understand and infer, then it should have nearly "equal" performances on the corpora of different statistical complexities since the understanding should be general. If there is a considerable decline in the performance, some reasons should be further explored and discussed.

[1]Yunyi Xu (260820529): yunyi.xu@mail.mcgill.ca
[2]Xiangyu Liu (260848712): xiang.yu.liu@mail.mcgill.ca
[3]Xinyuan Mo (260859175): xinyuan.mo@mail.mcgill.ca

III. Related Works

Paraphrase identification has been a widely studied NLP topic for many years [6][7][9]. Previous studies reported high accuracy in on including training Decomposable Attention (DECATT) Model using pre-trained word embeddings [11] and the bilateral multi-perspective matching (BiMPM) model [8]. Even though the two models showed high accuracy in the paraphrase identification task, both were trained on the Quora dataset. This dataset consists of sentence pairs with lower word overlap and simpler sentence structures compared to the PAWS, the dataset created based on Quora [13]. A recent study applied the Topic-Aware Paraphrase Identification Architecture (TAPA) to paraphrase identification on both Quora and PAWS, and it showed a considerable performance drop on PAWS compared with Quora [11]. Similar to the study by Peinelt et al [10], our study trained models on both Quora to PAWS, but with different language models. On the other hand, our aim is to compare the performance of language models on the two data sets as a measurement to test the preciseness of semantic parsing of the models.

IV. Experiment Setup

Model

We came up with 3 models with different complexities, namely Bag-of-n-grams (BONG), BERT, and MPNet.

BONG is a relatively simple model. We experimented with a variety of ranges of n. We noticed that to a certain extend, increasing n gives better performance. Other feature extraction techniques, such as stemming are used in a few experiments. Logistic regression is

naturally used as the classifier since the problem is a binary classification.

For BERT and MPNet, we adapted the state-of-the-art pretrained models (bert-base-nli-mean-tokens and all-mpnet-base-v2) by sentence-transformer[12] to get the sentence embeddings. Then for each sentences pair, we computed the cosine similarity between embeddings to measure the semantic similarity of two sentences. Using the label and cosine similarity of each pair, we trained a logistic regression model to make predictions for the test set.

Dataset	Sentence 1	Sentence 2	Label
Quora	Is USA the most powerful country of the world?	Why is the USA the most powerful country of the world?	0
	How can I be a good geologist?	What should I do to be a great geologist?	1
PAWS	In January 2011, FIBA Asia deputy secretary general Manuel V. Pangilinan along with SBP president Hagop Khajirian inspected the venue.	In January 2011, the Deputy Secretary General of FIBA Asia, Hagop Khajirian, inspected the venue together with SBP - President Manuel V. Pangilinan.	0
	On 6 March 2016, he debuted in the Ukrainian Premier League for FC Metalist Kharkiv in a game against FC Volyn Lutsk.	He made his debut in the Ukrainian Premier League for FC Metalist Kharkiv in a game against FC Volyn Lutsk on 6 March 2016.	1

Fig. 1: A few examples of both datasets.

Dataset	% Of labels = 0	% Of labels = 1
Quora	63.08	36.92
PAWS Train	55.80	44.20
PAWS Test	55.80	44.20

Fig. 2: Dataset statistics.

Dataset

We leveraged 2 datasets, namely the Quora Question Pairs (Quora) and the Paraphrase Adversaries from Word Scrambling (PAWS).

Both Quora and PAWS classify sentence pairs into 0 (not paraphrases) or 1 (paraphrases). Fig. 1 gives examples of entries of both datasets. Fig. 2 shows the ratio of the two classes of both datasets.

From Fig. 1, We can tell that the Quora dataset consists only of interrogative sentences, and they are generally short, more colloquial, and more rigid and simple in structures. Another feature of the Quora dataset is that the pairs with high lexical overlap are more likely to be paraphrases, which makes the statistical features of paraphrases more prominent.

Whereas the sentences from the PAWS dataset are longer, more formal, and more complex and flexible in structures, which requires stronger ability of understanding and infering for the paraphrase identification models.

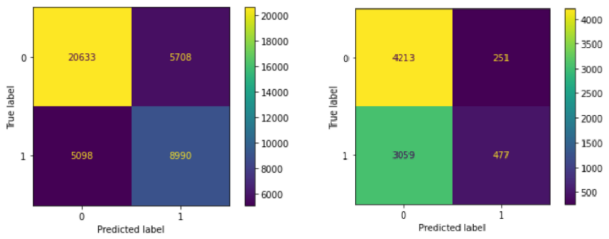

Fig. 3: Confusion Matrices with BERT on Quora(L) and PAWS(R)

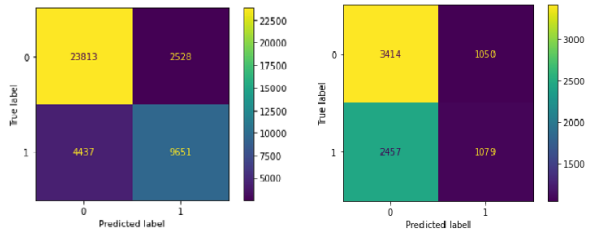

Fig. 4: Confusion Matrices with BONG on Quora(L) and PAWS(R)

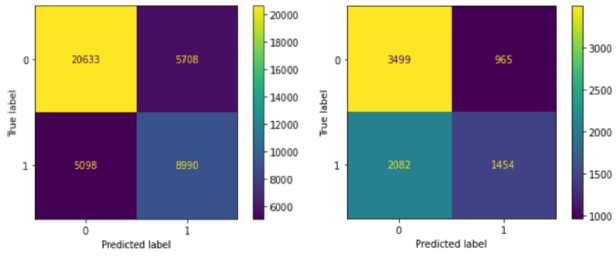

Fig. 5: Confusion Matrices with MPNET on Quora(L) and PAWS(R)

Model	Data Set	Accuracy	F1 score	Recall	Precision
MPNet	Quora	0.810	0.740	0.774	0.688
	PAWS	0.620	0.488	0.411	0.601
BERT	Quora	0.733	0.625	0.638	0.612
	PAWS	0.586	0.224	0.135	0.655
BONG	Quora	0.828	0.735	0.685	0.792
	PAWS	0.562	0.381	0.305	0.507

Fig. 6: Experiment Results

V. Results

The results can be seen in Figure.5. Overall, on the Quora Question Pairs dataset, we see that the BONG model obtains the best result, and on the PAWS dataset, the best result is achieved by the MPNet model. However, all three models have high performances

on the Quora Question Pairs dataset, while their performances all degenerate on the PAWS dataset.

The average accuracy of the three models on the Quora Dataset is around 0.79, while the average accuracy on the PAWS is around 0.60. The average F1 score of the three models on the Quora Dataset is around 0.70, while the average accuracy on the PAWS is only 0.37. Compared with the accuracy, the decline of the F1 score is more significant.

Surprisingly, the relatively simple BONG model achieves the best performance on the Quora Dataset. However, compared with the other two deep models, its performance also has the most severe descend on the PAWS dataset, in which many sentence pairs that have high lexical overlap are not paraphrases.

VI. Discussion

We have observed that the performances of all three models have declined significantly on the PAWS dataset, which suggests that the generalization ability of the deep networks models may not be so strong. In a dataset like PAWS, there are many sentence pairs that have high lexical overlap without being paraphrases, for example, the pair 'flights from New York to Florida' and 'flights from Florida to New York'. To address a suitable identification for such pairs, a model cannot only focus on the statistical features of the language forms; it requires the language model to capture and understand the semantic meanings and even to infer. Adapting to the modern NLP tasks, especially which require a deep understanding of the sentences and texts (like paraphrases identification), a language model based on distributional semantics (understanding a term by the distribution of words that appear near the term) may not achieve the desired result. And obviously, neither the simpler BONG model nor the sophisticated pretrained models performed well on this complex dataset. Notably, the decline of the BERT and MPNet is smaller than BONG, which is a sign that the BERT and MPNet indeed can extract the semantic meaning of a sentence to a certain extent, but it is not precise or powerful enough.

A limitation of the models is the difficulty in inferring certain concepts without external knowledge. For example, in our experiments, the models mislabel the pair "Where can rent PS4 game in Bengaluru?" and "Where can I rent a PS4 in Bangalore?" as non-paraphrases, whereas they are actually paraphrases of each other. Notice that "Bengaluru" and "Bangalore" refer to the same region. Similarly, the pair "Why Twitter CEO was not called at tech summit?" and "Why wasn't Jack Dorsey at the Trump Tech Summit?" were mislabelled as non-paraphrases by the models.

VII. Conclusion

We investigated to what extent does language models learn semantics, by solving paraphrase identification problems of different statistical complexities. We found that all our models perform significantly better with the easy dataset than the difficult one, therefore giving the conclusion that the models are not powerful enough to capture the semantic features and understand the meaning of the sentences.

VIII. Contributions

1) Yunyi Xu: Writing the Bag-of-n-grams model. Writing part I, IV, VI and VII of the report.
2) Xiangyu Liu: Writing the BERT and MPNET models. Writing part II, IV, V of the report.
3) Xinyuan Mo: Writing part I, III, and VI of the report.

References

[1] Google-Research-Datasets. (n.d.). Google-Research-datasets/PAWS: This dataset contains 108,463 human-labeled and 656K noisily labeled pairs that feature the importance of modeling structure, context, and word order information for the problem of paraphrase identification. GitHub. Retrieved December 21, 2021, from https://github.com/google-research-datasets/paws

[2] Bhagat, R., amp; Hovy, E. (2013). What is a paraphrase? Computational Linguistics, 39(3), 463–472. https://doi.org/10.1162/coli-a-00166

[3] Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[4] MPNet: Masked and permuted pre-training for language ... (n.d.). Retrieved December 21, 2021, from https://proceedings.neurips.cc/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf

[5] Bender, E. M., amp; Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

[6] Das, D., Smith, N. A. (2009). Paraphrase identification as probabilistic quasi-synchronous recognition. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - ACL-IJCNLP '09. https://doi.org/10.3115/1687878.1687944

[7] He, H., Gimpel, K., Lin, J. (2015). Multi-perspective sentence similarity modeling with Convolutional Neural Networks. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d15-1181

[8] Tomar, G. S., Duque, T., Täckström, O., Uszkoreit, J., Das, D. (2017). Neural paraphrase identification of questions with Noisy Pretraining. Proceedings of the First Workshop on Subword and Character Level Models in NLP. https://doi.org/10.18653/v1/w17-4121

[9] Triantafillou, E., Kiros, J. R., Urtasun, R., Zemel, R. (2016). Towards generalizable sentence embeddings. Proceedings of the 1st Workshop on Representation Learning for NLP. https://doi.org/10.18653/v1/w16-1628

[10] Peinelt, N., Nguyen, D., Liakata, M. (2020). Better early than late: fusing topics with word embeddings for neural question paraphrase identification. Arxiv, (2020 07 22).

[11] UKPLab. (n.d.). UKPLAB/sentence-transformers: Multilingual sentence amp; image embeddings with bert. GitHub. Retrieved December 21, 2021, from https://github.com/UKPLab/sentence-transformers

[12] Wang, Z., Hamza, W., amp; Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2017/579

[13] Zhang, Y., Baldridge, J., He, L. (2019). Paws: paraphrase adversaries from word scrambling. Arxiv, (2019 04 01)