Word-level Textual Adversarial Attacking as Combinatorial Optimization

1. **What is the primary lesson(s) you took away from this paper (avoid abstract level summary)?**

   The authors in this paper introduced a novel black-box word-level adversarial attack, SememePSO-Attack, that incorporates the sememe-based word substitution method and particle swarm optimization-based search algorithm. Firstly, the authors designed a word substitution method based on sememes (the minimum semantic units), which could retain more potential valid adversarial examples with high quality. Secondly, they presented a search algorithm based on particle swarm optimization, which is very efficient and performs better in finding adversarial examples.

   The authors performed exhaustive experiments to evaluate their black-box word-level adversarial attack model with automatic and human evaluations. They evaluated their attack with the baseline models (bidirectional-LSTM and BERT models), where their model attack not only achieved the highest attack success rate (e.g., 100% when attacking the bidirectional-LSTM on the IMDB movie reviews dataset) but also possessed the best adversarial example quality and comparable attack validity. Lastly, the authors demonstrated that their model has the highest transferability and could improve the robustness of the targeted models through adversarial training.

2. **Could I have done this work if I had the idea why or why not?**

   I do not think I could have done this work if I had the idea, given my limited knowledge of various model architectures, such as Recurrent Neural Networks (RNNs) like LSTM or bidirectional-LSTM, and BERT model (I am taking CS570 Deep Learning class this semester).

3. **Is there anything I could do to repeat or validate? (Make notes of materials available from the paper- data sets, source code).**

   Yes, the authors open-sourced their attack (SememePSO-Attack) and published it on GitHub: https://github.com/thunlp/SememePSO-Attack, where they shared their codes, models, and datasets.

   The authors also used three publicly available datasets: IMDB Movies Reviews dataset [1], Stanford Sentiment Treebank v2 (SST2) dataset [2], and the Stanford Natural Language Inference (SNLI) Corpus [3].

   [1] https://ai.stanford.edu/~amaas/data/sentiment/.
   [2] https://www.kaggle.com/datasets/atulanandjha/stanford-sentiment-treebank-v2-sst2.
   [3] https://nlp.stanford.edu/projects/snli/.

4. **What is my best idea for follow on work that I could personally do?**

   My best idea for the follow-on work that I could personally do is to first repeat the same work using their English datasets to develop a better understanding of how the attack works. Secondly, I could repeat the same work or part of using Arabic text classification datasets, examine the generated results, and test if this kind of adversarial text attack works in different languages like Arabic.

5. **What is my best idea for follow on work that I'd like to see the authors or others do?**

   My best idea for the follow-on work that I would like to see the authors or others do is to "*try to increase the robustness gains of adversarial training and consider utilizing sememes in adversarial defense model.*"—the authors; they have proposed a word-level attack model comprising the sememe-based word substitution method but haven't used it in the adversarial training as they stated.

6. **Any logistical experimental lessons I learned (this is a little different than #1 here you are looking for techniques)?**

   I liked how the authors provided an end-to-end evaluation in their paper, covering almost every evaluation metric discussed in the research literature. I also liked the idea of dividing the evaluations into two categories: automatic evaluation metrics (success rate, modification percentage, and fluency) and human evaluations (validity and naturality).

   Another experimental lesson I learned from his paper is how the authors conducted detailed decomposition analyses of different word substitution methods (search space reduction methods) and different search algorithms to demonstrate further the advantages of their sememe-based word substitution method and PSO-based search algorithm.

7. **How does this compare to the other papers we read? Most similar? How different? Other comparisons?**

   Zang et al. (2020) is the first paper that uses the sememe-based word substitution method, yet it is among ten papers I have read in this course (directed study) that discussed the black-box attacks of DNN models. In terms of the attack type (black-box attacks), this paper is similar to ten papers and different from four papers (white-box attacks); in terms of the space type (input space), this paper is similar to sixteen papers and different from four papers (embedding space); and lastly, in terms of the method used (word-level replacements), this paper is similar to thirteen papers and different from seven papers (character-level replacements and others).

8. **What is your biggest criticism of the paper?**

   There is *no* big criticism of the paper.

9. **List 3 cited references or terms/concepts that you would be most interested in reading/learning more about.**

   I am interested in reading these three references because they are very relevant, and the authors of this paper followed some of their approaches.

   Xiaosen Wang, Hao Jin, and Kun He. 2019b. ***Natural language adversarial attacks and defenses in word level****. arXiv preprint arXiv:1909.06723.*

   Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. ***Adversarial examples in the physical world.*** *arXiv preprint arXiv:1607.02533.*

   Russell Eberhart and James Kennedy. 1995. ***Particle swarm optimization****. In Proceedings of the IEEE International Conference on Neural Networks.*