Combating Adversarial Misspellings with Robust Word Recognition

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

The authors in this paper proposed a defense mechanism to battle adversarial misspellings with robust word recognition models, semi-character RNNs (ScRNN), that predict each word in a sentence given a full sequence of possibly misspelled inputs, where these word recognition models' outputs comprise the input to a downstream classification model. Since these word recognition models are trained on domain-specific text, and as a result, these models often predict unknown words at test time because of the small domain-specific vocabulary. The authors also proposed several *backoff* strategies (pass-through, backoff to a neutral word like 'a', or backoff to a background model to correct the misspellings) to manage the unobserved and rare words, including falling back on a word recognizer (background model) trained on a larger corpus. These backoff strategies, when incorporated with defense mechanisms, BERT models, for example, subject to one-character attacks, are restored to 88.3%, 81.1%, and 78.0% accuracy rates for swap, drop, and add attacks, respectively, as compared to the restored accuracy rates using adversarial training 69.2%, 63.6%, and 50.0%.

The authors in experiments used bidirectional-LSTMs and fine-tuned BERT models comprising four different input formats: word-only, character-only (char-only), word+character (word+char), and word-piece. This requires altering just two characters per sentence, where such modifications might flip words to a different word in the vocabulary or, more often, to the out-of-vocabulary (OOV) as an unknown token (UNK). They showed in their experiments that character and word-piece models are in fact *more vulnerable* than word-only models. In the word-level models case, the adversary is mostly limited to UNK words, whereas in the case of the word-piece or character-level models, each character-level add, drop, or swap produces a distinct input, providing the adversary with a greater set of options.

The authors evaluated the first-line techniques, including data augmentation and adversarial training, demonstrating that these techniques offer only marginal robustness. For example, a BERT model achieving 90.3% accuracy on a sentiment classification task can be degraded to 64.1% by an adversarially-chosen one-character swap in the sentence, which can only be restored to 69.2% by adversarial training. Lastly, the authors showed a detailed qualitative analysis to demonstrate that a low word error rate (WER) alone is insufficient for a word recognizer model to confer robustness on the downstream task. Therefore, the authors provided a metric to quantify this notion of *sensitivity* in word recognition models and empirically study its relation to robustness, where models with low sensitivity and word error rate are the most robust.

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 RNNs, bidirectional-LSTMs, and BERT (I am taking CS570 Deep Learning class this semester). I also do not think I could mathematically define the notion of sensitivity in the word recognition models, even if I had the idea.

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 are open-sourcing their defenses, attacks, and baselines, and they are all available at GitHub here https://github.com/danishpruthi/Adversarial-Misspellings. The authors also benchmarked their semi-character RNNs (ScRNN) models with an open-source spell corrector, *After The Deadline* (ATD), which is available at GitHub here https://github.com/Automattic/atd-core.

4. **What is my best idea for follow on work that I could personally do?**
   I do not have an idea for the follow-on work that I could personally do.

5. **What is my best idea for follow on work that I'd like to see the authors or others do?**
   I do not have an idea for the follow-on work that I would like to see the authors or others do.

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 evaluated the first-line techniques used to confer robustness to NLP models: data augmentation and adversarial training. They demonstrated that these two techniques can only offer marginal robustness.

   I also liked the idea of implementing an attack based on substituting an internal character with adjacent characters of the QWERTY keyboard to mimic people's unintentional misspellings/typos on tiny screens like smartphones.

7. **How does this compare to the other papers we read? Most similar? How different? Other comparisons?**
   This paper, Pruthi et al. (2019), is the second paper I have read about defense mechanisms against adversarial attacks, especially back-box attacks where the attackers are unaware of the model architecture, parameters, or training datasets. This paper is similar to Alshemali & Kalita (2019) in terms of defense type -- correcting misspellings and input space, yet they are different in the method used. Alshemali & Kalita (2019) proposed a novel approach of spell-checking systems that utilize frequency and contextual information for correcting nonword misspellings, while this paper, Pruthi et al. (2019), proposed robust word recognition models, semi-character RNNs (ScRNN), to battle adversarial misspellings.

8. **What is your biggest criticism of the paper?**
   There is no big criticism of the paper. I liked the idea of summarizing the key results besides their corresponding intros in the Introduction section. Yet, this idea could keep readers away from the rest of the paper and its detailed experiments and results.

9. **List 3 cited references or terms/concepts that you would be most interested in reading/learning more about.**
   I am interested in reading this reference, Sakaguchi et al. (2017), because the authors built their word recognition models upon the RNN-based semi-character word recognition models that were initially introduced in this reference.

   Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. ***Robsut Wrod Reocginiton via semi-Character Recurrent Neural Network****. In Association for the Advancement of Artificial Intelligence (AAAI).*

   I am also interested in reading these two references (Krizhevsky et al., 2012 and Goodfellow et al., 2014) that discuss baseline defense strategies; the two common methods for dealing with adversarial examples are data augmentation and adversarial training.

   Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ***Imagenet classification with deep convolutional neural networks.*** *In Advances in neural information processing systems (NIPS).*

   Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. ***Explaining and harnessing adversarial examples.*** *In International Conference on Learning Representations (ICLR).*