Progress Report of November 2022

This past month of November 2022, I have read seven research papers: five papers published in 2020 and two published in 2021. Two of these papers discussed frameworks, which are TextAttack and CheckList. TextAttack is a framework for adversarial attacks, data augmentation, and adversarial training in NLP, whereas CheckList is a framework for behavioral testing of NLP models.

The new in this month's reading list is that it converged new topics in this domain, such as discussing character-level black-box attacks in the Arabic language, discussing sub-word-level BERT-based adversarial examples, and discussing contextualized perturbation on DNNs. Here is a list of these papers:

1. Jin et al. *Is Bert Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment*. AAAI 2020.
2. Morris et al. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP.* EMNLP 2020.
3. Li et al. *BERT-ATTACK: Adversarial Attack Against BERT Using BERT*. EMNLP 2020.
4. Ribeiro et al. *Beyond Accuracy: Behavioral Testing of NLP models with CheckList.* ACL 2020.
5. Zang et al. *Word-level Textual Adversarial Attacking as Combinatorial Optimization.* ACL 2020.
6. Alshemali and Kalita. *Character-level Adversarial Examples in Arabic.* ICMLA 2021.
7. Li et al. *Contextualized Perturbation for Textual Adversarial Attack.* NAACL 2021.

Throughout my reading of these research papers, I kept a brief taxonomy of the adversarial examples/attacks in terms of the attack type, the targeted space, and the method used, hoping to turn this taxonomy into a review paper. The table below shows the taxonomy of the 2020 and 2021 papers.

#	Paper	Authors	Attack	Space	Method
1	Is Bert Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment.	Jin et al. (2020)	Black-box Attacks	Input	Word-level Replacements
2	TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP	Morris et al. (2020)	Framework	Input	*Word-level Replacements **Character-level Replacements
3	BERT-ATTACK: Adversarial Attack Against BERT Using BERT	Li et al. (2020)	Black-box Attacks	Input	*Word-level Replacements **Sub-word-level Replacements
4	Beyond Accuracy: Behavioral Testing of NLP models with CheckList	Ribeiro et al. (2020)	Framework	Input	*Word-level Replacements **Character-level Replacements
5	Word-level Textual Adversarial Attacking as Combinatorial Optimization	Zang et al. (2020)	Black-box Attacks	Input	Word-level Replacements
6	Character-level Adversarial Examples in Arabic	Alshemali and Kalita (2021)	Black-box Attacks	Input	Character-level Replacements
7	Contextualized Perturbation for Textual Adversarial Attack	Li et al. (2021)	Black-box Attacks	Input	Word-level Replacements

--- End of Report ---