Appendix

A Hyper-parameters

The hyper-parameters of the MHA model, the genetic baseline model, and the victim models are listed as follow.

MHA. We set the hyper-parameters of MHA to $p_r = 0.5$, $p_i = 0.25$, $p_d = 0.25$. Constraints on $LM(x)$ and $C(\tilde{y}|x)$ is performed – if $LM(x') < t_{LM} \cdot LM(x)$ or $C(\tilde{y}|x') < t_C \cdot C(\tilde{y}|x)$, the proposal is rejected directly. Such trick ensures that we do not loss sentence fluency or target probability rapidly. t_{LM} and t_C are set to 0.8 and 0.9 in our experiments. Also, any operation on sentimental words (eg. "great") or negation words (eg. "not") are forbidden in IMDB experiments. SentiWordNet (Esuli and Sebastiani, 2006; Baccianella et al., 2010) are applied to recognize the sentimental words.

The language models in MHA includes a forward and a backward 2-layer LSTM models with 300 units trained on subset of the One-Billion-Word Corpus (Chelba et al., 2013). We randomly select 5M sentences from the corpus for LM training. The vocabulary size is 50,000. The two LSTMs employ independent embedding matrices with the same word2vec initialization.

Genetic baseline. The hyper-parameter settings of the genetic model remain the same as in the paper (Alzantot et al., 2018).

Victim models. The LSTMs in the victim models have 128 units. The bi-LSTM for IMDB has a vocabulary size of 10,000, while the two LSTMs in the BiDAF model share the same vocabulary size of 35,000. The embedding matrices are pretrained by word2vec. In addition, the embedding matrix of the bi-LSTM model for IMDB is fixed during training to avoid overfitting. All classifiers in our experiment reach 99% accuracy on the training set.

B Adversarial Examples

Some adversarial examples are listed in Table 1. The genetic replacement considers only the current word itself, regardless of its context, and results in an unfluent sentence, while MHA performs replacement with the guide of LM, and the sentence is fluent.

Empirically, MHA often operates the prepositions, the pronouns, and the punctuations, *etc.*, where the changes are minor, while the genetic approach only replace the verbs, the nouns, the adjectives and the adverbs. An advantage of operating the prepositions, the punctuations, *etc.* is that human beings usually do not pay much attention to them. Human begins can hardly recognize the adversarial examples generated by these operations.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer.

Examples
Premise: *a group of people examine a boat with an orange flag that is sitting on sand next to a body of water.* **Original hypothesis:** *people sit on a beach to tan.* **Prediction:** ⟨Neutral⟩ **Genetic hypothesis:** *people sit on a* ***swimming*** *to tan.* **Prediction:** ⟨Contradiction⟩ **_b_-MHA hypothesis:** *people sit on a beach* ***and*** *tan.* **Prediction:** ⟨Contradiction⟩ **_w_-MHA hypothesis:** *people sit on a beach* ***and*** *tan.* **Prediction:** ⟨Contradiction⟩
Premise: *a woman lying in the grass in the park is wearing a red top and black capri pants and is barefoot.* **Original hypothesis:** *a woman is sitting on a park bench wearing sandals.* **Prediction:** ⟨Contradiction⟩ **Genetic hypothesis:** *a woman is* ***seated*** *on a park bench wearing* ***footwear****.* **Prediction:** ⟨Entailment⟩ **_b_-MHA hypothesis:** *a woman is sitting on a park bench wearing* ***it****.* **Prediction:** ⟨Entailment⟩ **_w_-MHA hypothesis:** *a woman is sitting on a park bench wearing* ***it****.* **Prediction:** ⟨Entailment⟩
Premise: *a man alone crosscountry skis in the wilderness while wearing a huge backpack.* **Original hypothesis:** *a man skis in the wilderness while it's snowing* **Prediction:** ⟨Neutral⟩ **Genetic hypothesis:** *a man* ***snowboarding*** *in the wilderness while it's snowing* **Prediction:** ⟨Contradiction⟩ **_b_-MHA hypothesis:** *a man skis in the* ***world*** *while it's snowing* **Prediction:** ⟨Contradiction⟩ **_w_-MHA hypothesis:** *a man skis in the wilderness* ***and*** *it's snowing* **Prediction:** ⟨Contradiction⟩
Premise: *a boy kneeling on a skateboard riding down the street* **Original hypothesis:** *a boy standing upright on a skateboard.* **Prediction:** ⟨Contradiction⟩ **Genetic hypothesis:** *a boy* ***permanent*** *upright on a skateboard.* **Prediction:** ⟨Entailment⟩ **_b_-MHA hypothesis:** *a boy* ***is out*** *on a skateboard .* **Prediction:** ⟨Entailment⟩ **_w_-MHA hypothesis:** *a boy* ***was*** *upright on a skateboard.* **Prediction:** ⟨Entailment⟩
Premise: *three men are sitting on a beach dressed in orange with refuse carts in front of them.* **Original hypothesis:** *empty trash cans are sitting on a beach.* **Prediction:** ⟨Contradiction⟩ **Genetic hypothesis:** ***empties*** *trash cans are sitting on a beach..* **Prediction:** ⟨Entailment⟩ **_b_-MHA hypothesis:** ***the*** *trash cans are sitting* ***in*** *a beach.* **Prediction:** ⟨Entailment⟩ **_w_-MHA hypothesis:** ***the*** *trash cans are sitting on a beach.* **Prediction:** ⟨Entailment⟩
Premise: *hikers walk along some tough terrain.* **Original hypothesis:** *hiking pace along rough terrain.* **Prediction:** ⟨Entailment⟩ **Genetic hypothesis:** *hiking pace along rough* ***terra****.* **Prediction:** ⟨Neutral⟩ **_b_-MHA hypothesis:** *hiking* ***is in*** *rough terrain.* **Prediction:** ⟨Neutral⟩ **_w_-MHA hypothesis:** ***the*** *pace along rough terrain.* **Prediction:** ⟨Neutral⟩
Premise: *our people walking beside each other down a street, one of the men is turned around looking toward the camera.* **Original hypothesis:** *a group of friends are headed to wendys* **Prediction:** ⟨Neutral⟩ **Genetic hypothesis:** *a groups of* ***boyfriends*** *are* ***guided*** *to wendys* **Prediction:** ⟨Contradiction⟩ **_b_-MHA hypothesis:** *a group of* ***women*** *are* ***expected*** *to wendys* **Prediction:** ⟨Contradiction⟩ **_w_-MHA hypothesis:** *a* ***number*** *of* ***people*** *are* ***going*** *to wendys* **Prediction:** ⟨Contradiction⟩
Premise: *a man in a green shirt hovers above the ground in the laundry room.* **Original hypothesis:** *the man appears to be suspended in midair.* **Prediction:** ⟨Entailment⟩ **Genetic hypothesis:** *the man* ***emerge*** *to be suspended in midair.* **Prediction:** ⟨Neutral⟩ **_b_-MHA hypothesis:** *the man appears to be suspended in* ***2007****.* **Prediction:** ⟨Neutral⟩ **_w_-MHA hypothesis:** *the man* ***is*** *to be suspended in midair.* **Prediction:** ⟨Neutral⟩

Table 1: Adversarial examples generated on SNLI.