1. Summary of One-level Black-box Adversarial Applications

Adversary	Perturbation Methods	Perturbation Level	Architectures	Dataset
Heigold et al. [23].	1. Randomly swap two adjacent characters in a word. 2. Randomize the order of all the characters in a word except for the first and the last. 3. Randomly replace a character with another character at a pre-specified rate.	Character-level.	CNN, RNN.	UD English dataset, German-English (DEEN) parallel corpora provided by WMT16.
Belinkov and Bisk [4].	1. Randomly swap two adjacent characters in a word. 2. Randomize the order of all the characters in a word except for the first and the last. 3. Randomize the order of all the characters in a word. 4. Randomly replace a single character with the character next to it on the English keyboard.	Character-level.	char2char [28], Nematus [45], charCNN [54].	TED talks (French, German, and Czech) to English parallel corpus.

Gao et al. [19].	1. Randomly swap two adjacent characters in a word. 2. Randomly replace a character with another character. 3. Randomly delete a character. 4. Randomly insert a character.	Character-level.	Word-LSTM, charCNN [54].	AG's News, Amazon Review full and polarity, DBPedia, Yahoo! Answers, Yelp Review full and polarity, Enron Spam.
Sogaard et al. [49].	1. Delete all punctuation marks. 2. Insert commas and dots.	Character-level.	UUPARSER [12] [13], KGRAPHS [27], STANFORD [6], MALTPARSER [39], TURBOPARSER [18].	English Penn Treebank, Google Web Treebank.
Samanta and Mehta [44].	1. Replace adjectives with adverbs. 2. Insert adverbs. 3. Remove adverbs.	Token-level.	CNN.	Twitter dataset for gender classification, IMDB.
Alzantot et al. [1].	1. Replace a token with a semantically similar token.	Token-level.	LSTM.	IMDB.
Glockner et al. [20].	1. Replace a token with its synonym or hypernym. 2. Replace a token with its hyponym or antonym.	Token-level.	ESIM [9], Decomposable Attention [40], Residual-Stacked-Encoder [37], WordNet [17], KIM [7].	SNLI, MultiNLI, SciTail.

Jia and Liang [26].	1. ADD-SENT: Insert grammatical sentences that are similar to the question but do not conflict with the correct answer. 2. ADD-ANY: Insert arbitrary sequences of English words, regardless of grammaticality. 3. ADD-COMMON: Like ADD-ANY except that it only Inserts English common words. 4. ADD-ONE-SENT: Insert a human-approved sentence, selected at random.	Sentence-level.	jNet [53], BiDAF Single and Ensemble versions [46], RaSOR [29], Match-LSTM Single and Ensemble versions [51], Ruminating Reader [21], Logistic Regression Baseline [41], Dynamic Chunk Reader (DCR)[1], ReasoNet Single and Ensemble versions [48], Mnemonic Reader Single and Ensemble versions [25], Multi-Perspective Context Matching (MPCM) Single version[2], Structural Embedding of Dependency Trees (SEDT) Single and Ensemble versions [32].	SQuAD.
Ribeiro et al. [42].	1. Paraphrase sentences.	Sentence-level.	FastText, Zhu et al.'s [55].	Rotten Tomatoes movie reviews, IMDB sentence-sized reviews, Zhu et al.'s [55].

Table 1: A summary of the black-box adversarial applications that make perturbations on one level: Character-level, Token-level, or Sentence-level.

[1] arXiv:1610.09996
[2] arXiv:1612.04211

2. Summary of Two-level Black-box Adversarial Applications

Adversary	Perturbation Methods	Perturbation Level	Architectures	Dataset
Naik et al. [35].	1. Change numbers or prefix them with “less than” or “more than”. 2. Replace tokens with their antonyms. 3. Append the tautology “and true is true” to the end of every hypothesis sentence (word overlapping). 4. Append the tautology “and false is not true” to the end of every hypothesis sentence (negation). 5. Append the tautology “and true is true” five times to the end of every premise sentence (length mismatching). 6. Randomly swap adjacent characters within a word. 7. Randomly replace a single character with the character next to it on the English keyboard.	Character-level, token-level.	Nie and Bansal’s model [37], Chen et al.’s model [8], Balazs et al.’s model [2], Conneau et al.’s model [10], BiLSTM [36], CBOW [33].	MultiNLI.
Blohm et al. [5].	1. Replace the most frequent question words with manually defined meaning-preserving lexical substitutions. 2. Insert a distracting sentence contains random words from common English words (AddC). 3. Insert a distracting sentence contains words from the question words (AddQ) 4. Insert a distracting sentence contains words from the question and incorrect answers (AddQA).	Token-level, sentence-level.	Wang and Jiang’s model [50], Liu et al.’s model, Dzendzik et al.’s model [14], CNN word level, CNN, CNN ensemble, RNN-LSTM, RNN-LSTM ensemble, CNN RNN-LSTM ensemble.	MovieQA.

Niu and Bansal [38].	1. Randomly swap two adjacent tokens. 2. Randomly delete stop-words. 3. Replace tokens with a paraphrasing. 4. Replaces grammatically correct words or phrases with wrong ones. 5. Paraphrase sentences. 6. Negate verbs. 7. Replace verbs, adverbs, or adjectives with their antonyms.	Token-level, sentence-level.	VHRED [47], RL, DynoNet [22].	CoCoA, Ubuntu Dialogue Corpus.
Henderson et al. [24].	1. Misspelling words by removing, replacing or inserting an extra character in the word. 2. Paraphrasing sentences.	Character-level, sentence-level.	VHRED.	Reddit Movies, Reddit Politics.

Table 2: A summary of the black-box adversarial applications that make perturbations on two levels: Character-token-level, Token-sentence-level, or Character-sentence-level.

3. Summary of White-box Adversarial Applications

Adversary	Perturbation Methods	Perturbation Scope	Architectures	Dataset
Behjati et al. [3].	1. Insert a token or a sequence of tokens.	Token-level.	LSTM, bi-LSTM, mean-LSTM.	AG's news, SST.
Mudrakarta et al. [34].	1. Insert grammatical sentences that include important tokens from the questions.	Sentence-level.	Yu et al.'s model [52].	SQuAD.
Ebrahimi et al. [16].	1. Character or token flipping. 2. Character or token inserting. 3. Character or token removal.	Character-level, token-level.	charCNN-LSTM, Word-CNN.	AG's news, SST.
Ebrahimi et al. [15].	1. Replacing a character with another. 2. Swapping two adjacent characters. 3. Deleting a character. 4. Inserting a character. 5. Removing a specific token from the translation output. 6. Replace a token with another in the translation output.	Character-level, token-level.	Costa-Jussa et al.'s model [11].	TED talks (French, German, and Czech) to English parallel corpora.

Blohm et al. [5].	1. Substitute the words that receive most attention with randomly chosen words. 2. Remove the sentence with the highest attention.	Token-level, sentence-level.	Wang and Jiang's model [50], Liu et al.'s model, Dzendzik et al.'s model [14], CNN word level, CNN, CNN ensemble, RNN-LSTM, RNN-LSTM ensemble, CNN RNN-LSTM ensemble.	MovieQA.
Liang et al. [31].	1. Insert a token, phrase, or a sentence. 2. Replace a token with a misspelled version of it. 3. Replace a character with a character that has a similar visual appearance. 4. Delete a token, phrase, or a sentence.	Character-level, token-level, sentence level.	charCNN [54].	DBpedia.

Table 3: A summary of the white-box adversarial applications that make perturbations on: Token-level, Sentence-level, Character-token-level, Token-sentence-level, or Character-token-sentence level.

4. Summary of Compromised Real-world Applications

Adversary	Perturbation Methods	Perturbation Level	Applications	Dataset
Rodriguez and Rojas-Galeano [43].	1. Obfuscation: Replace characters, repeat characters and insert unnecessary punctuation marks within characters in the words (commas, dots, or blanks). 2. Polarity: Negate toxic words.	Character-level, token-level.	Google's Perspective API.	Google's Perspective dataset.
Li et al. [30].	1. Insert a space to the word; 2. Randomly delete a character from the word except for the first and the last character; 3. Randomly swap two adjacent characters in the word except for the first and the last character; 4. Replace characters with visually similar characters or with adjacent characters in the keyboard; 5. Replace a word with a semantically similar word.	Character-level, token-level.	Google Machine Learning, Microsoft Azure, IBM Watson, Facebook fast-Text, Amazon Machine Learning, ParallelDots, Aylien Sentiment, TheySay Sentiment, TextProcessing, Mashape Sentiment, Google Perspective.	MDB, Rotten Tomatoes movie reviews, the Kaggle Toxic Comment Classification competition dataset.

Table 4: A summary of the real-world applications that have been compromised by adversarial examples.

References

[1] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, K.-W. Chang, Generating natural language adversarial examples, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2890–2896.

[2] J. A. Balazs, E. Marrese-Taylor, P. Loyola, Y. Matsuo, Refining raw sentence representations for textual entailment recognition via attention, in: Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, 2017, pp. 51–55.

[3] M. Behjati, S.-M. Moosavi-Dezfooli, M. S. Baghshah, P. Frossard, Universal adversarial attacks on text classifiers, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7345–7349.

[4] Y. Belinkov, Y. Bisk, Synthetic and natural noise both break neural machine translation, in: International Conference on Learning Representations, 2018.
URL https://openreview.net/forum?id=BJ8vJebC-

[5] M. Blohm, G. Jagfeld, E. Sood, X. Yu, N. T. Vu, Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension, in: Proceedings of the 22nd Conference on Computational Natural Language Learning, 2018, pp. 108–118.

[6] D. Chen, C. Manning, A fast and accurate dependency parser using neural networks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 740–750.

[7] Q. Chen, X. Zhu, Z.-H. Ling, D. Inkpen, S. Wei, Neural natural language inference models enhanced with external knowledge, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2018, pp. 2406–2417.

[8] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, D. Inkpen, Recurrent neural network-based sentence encoder with gated attention for natural language inference, in: Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, 2017, pp. 36–40.

[9] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, D. Inkpen, Enhanced LSTM for natural language inference, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1657–1668.

[10] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017, pp. 670–680.

[11] M. R. Costa-Jussà, C. España-Bonet, P. Madhyastha, C. Escolano, J. A. Fonollosa, The TALP–UPC Spanish–English WMT biomedical task: Bilingual embeddings and char-based neural language model rescoring in a phrase-based system, in: Proceedings of the First Conference on Machine Translation, Vol. 2, 2016, pp. 463–468.

[12] M. de Lhoneux, Y. Shao, A. Basirat, E. Kiperwasser, S. Stymne, Y. Goldberg, J. Nivre, From raw text to universal dependencies-look, no tags!, Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (2017a) 207–217.

[13] M. de Lhoneux, S. Stymne, J. Nivre, Arc-hybrid non-projective dependency parsing with a static-dynamic oracle, in: Proceedings of the 15th International Conference on Parsing Technologies, 2017b, pp. 99–104.

[14] D. Dzendzik, C. Vogel, Q. Liu, Who framed Roger Rabbit? multiple choice questions answering about movie plot, in: Proceedings of the The Joint Video and Language Understanding Workshop: MovieQA and The Large Scale Movie Description Challenge (LSMDC), 2017.

[15] J. Ebrahimi, D. Lowd, D. Dou, On adversarial examples for character-level neural machine translation, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 653–663.

[16] J. Ebrahimi, A. Rao, D. Lowd, D. Dou, Hotflip: White-box adversarial examples for NLP, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 31–36.

[17] C. Fellbaum, Wordnet: Wiley online library, The Encyclopedia of Applied Linguistics.

[18] D. Fernández-González, A. F. Martins, Parsing as reduction, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, pp. 1523–1533.

[19] J. Gao, J. Lanchantin, M. L. Soffa, Y. Qi, Black-box generation of adversarial text sequences to evade deep learning classifiers, in: IEEE Security and Privacy Workshops, 2018, pp. 50–56.

[20] M. Glockner, V. Shwartz, Y. Goldberg, Breaking NLI systems with sentences that require simple lexical inferences, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 650–655.

[21] Y. Gong, S. R. Bowman, Ruminating reader: Reasoning with gated multi-hop attention, in: Proceedings of the Workshop on Machine Reading for Question Answering, 2018, pp. 1–11.

[22] H. He, A. Balakrishnan, M. Eric, P. Liang, Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1766–1776.

[23] G. Heigold, G. Neumann, J. van Genabith, How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse?, in:

Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, 2018, pp. 68–80.

[24] P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, J. Pineau, Ethical challenges in data-driven dialogue systems, in: AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2018, pp. 123–129.

[25] M. Hu, Y. Peng, Z. Huang, X. Qiu, F. Wei, M. Zhou, Reinforced mnemonic reader for machine reading comprehension, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 4099–4106.

[26] R. Jia, P. Liang, Adversarial examples for evaluating reading comprehension systems, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2021–2031.

[27] E. Kiperwasser, Y. Goldberg, Simple and accurate dependency parsing using bidirectional LSTM feature representations, in: Transactions of the Association for Computational Linguistics, Volume 4, Issue 1, 2016.

[28] J. Lee, K. Cho, T. Hofmann, Fully character-level neural machine translation without explicit segmentation, Transactions of the Association for Computational Linguistics 5 (2017) 365–378.

[29] K. Lee, S. Salant, T. Kwiatkowski, A. Parikh, D. Das, J. Berant, Learning recurrent span representations for extractive question answering, in: International Conference on Learning Representations, 2017.
URL `https://openreview.net/forum?id=HkIQH7qel`

[30] J. Li, S. Ji, T. Du, B. Li, T. Wang, TEXTBUGGER: Generating adversarial text against real-world applications, in: Proceedings of Network and Distributed System Security Symposium (NDSS), 2019.

[31] B. Liang, H. Li, M. Su, P. Bian, X. Li, W. Shi, Deep text classification can be fooled, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-18), 2017.

[32] R. Liu, J. Hu, W. Wei, Z. Yang, E. Nyberg, Structural embedding of syntactic trees for machine comprehension, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017, pp. 815–824.

[33] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013.
URL https://openreview.net/forum?id=idpCdOWtqXd60s

[34] P. K. Mudrakarta, A. Taly, M. Sundararajan, K. Dhamdhere, Did the model understand the question?, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 1896–1906.

[35] A. Naik, A. Ravichander, N. Sadeh, C. Rose, G. Neubig, Stress test evaluation for natural language inference, in: Proceedings of the International Conference on Computational Linguistics, 2018, pp. 2340–2353.

[36] N. Nangia, A. Williams, A. Lazaridou, S. R. Bowman, The Repeval 2017 shared task: Multi-genre natural language inference with sentence representations, in: Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, 2017, pp. 1–10.

[37] Y. Nie, M. Bansal, Shortcut-stacked sentence encoders for multi-domain inference, in: Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, 2017, pp. 41–45.

[38] T. Niu, M. Bansal, Adversariasl over-sensitivity and over-stability strategies for dialogue models, in: Proceedings of the 22nd Conference on Computational Natural Language Learning, 2018, pp. 486–496.

[39] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, E. Marsi, Maltparser: A language-independent system for data-driven dependency parsing, Natural Language Engineering 13 (2) (2007) 95–135.

[40] A. P. Parikh, O. Täckström, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2249–2255.

[41] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392.

[42] M. T. Ribeiro, S. Singh, C. Guestrin, Semantically equivalent adversarial rules for debugging nlp models, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2018, pp. 856–865.

[43] N. Rodriguez, S. Rojas-Galeano, Fighting adversarial attacks on online abusive language moderation, Applied Computer Sciences in Engineering 915 (2018) 480–493.

[44] S. Samanta, S. Mehta, Generating adversarial text samples, Advances in Information Retrieval 10772 (2017) 744–749.

[45] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. M. Barone, J. Mokry, et al., Nematus: a toolkit for neural machine translation, in: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 65–68.

[46] M. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, Bidirectional attention flow for machine comprehension, in: International Conference on Learning Representations, 2017.
URL `https://openreview.net/forum?id=HJ0UKP9ge`

[47] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, Y. Bengio, A hierarchical latent variable encoder-decoder model for generating dialogues., in: The Association for the Advancement of Artificial Intelligence, 2017, pp. 3295–3301.

[48] Y. Shen, P.-S. Huang, J. Gao, W. Chen, Reasonet: Learning to stop reading in machine comprehension, in: Proceedings of the 23rd ACM International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1047–1055.

[49] A. Søgaard, M. de Lhoneux, I. Augenstein, Nightmare at test time: How punctuation prevents parsers from generalizing, in: Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, pp. 25–29.

[50] S. Wang, J. Jiang, A compare-aggregate model for matching text sequences, in: International Conference on Learning Representations, 2017.
URL https://openreview.net/forum?id=HJTzHtqee

[51] S. Wang, J. Jiang, Machine comprehension using match-LSTM and answer pointer, in: International Conference on Learning Representations, 2017.
URL https://openreview.net/pdf?id=B1-q5Pqxl

[52] A. W. Yu, D. Dohan, Q. Le, T. Luong, R. Zhao, K. Chen, Fast and accurate reading comprehension by combining self-attention and convolution, in: International Conference on Learning Representations, 2018.
URL https://openreview.net/forum?id=B14TlG-RW

[53] J. Zhang, X. Zhu, Q. Chen, L. Dai, S. Wei, H. Jiang, Exploring question understanding and adaptation in neural-network-based question answering, in: Proceedings of the 3rd IEEE International Conference on Computer and Communications, 2017, pp. 1975–1984.

[54] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: Advances in Neural Information Processing Systems, 2015, pp. 649–657.

[55] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: Grounded question answering in images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4995–5004.