Philipps Universität Marburg

Controllable Generation Using kNN Language Models

Master/Bachelor Thesis

Motivation

Language modeling forms the foundation of many language processing problems and involves predicting which words come next in a sequence. Recent work on kNN language models has shown improved perplexity by storing encodings of sentence contexts and retrieving similar contexts to alter the probability distribution when predicting the next token (Khandelwal et al. 2020). If we store not only the encoding, but also labeled stylistic attributes of a sentence (e.g. politeness, formality, toxicity), can we use this to control generated language to contain these attributes to higher or lower degrees while preserving fluency? How can these stored attributes and encodings be leveraged most effectively?

Difficulty

Analysis

Programming

Literature

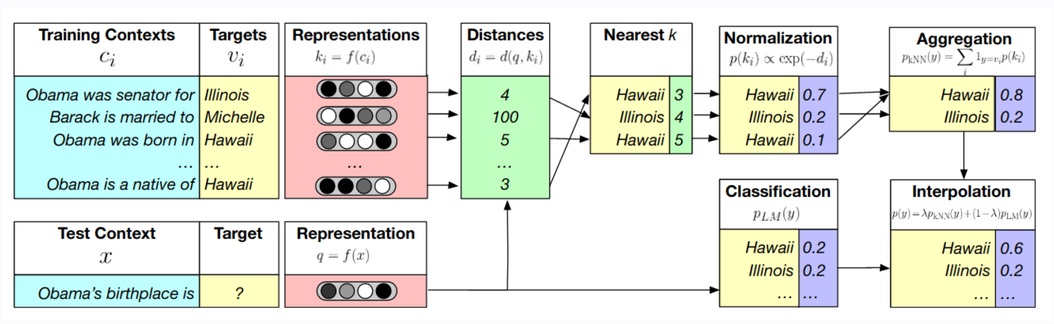
Contact

Prof. Dr. Lucie Flek
Dr. Charlie Welch

lucie.flek@gmail.com
cwelch@mathematik.uni-marburg.de
caisa-lab.github.io
06421 / 28 - 21586
06421 / 28 - 28902
CAISA Lab FB 12 | Mathematik und Informatik
Hans-Meerwein-Straße 6, 35043, Marburg
Room: 04C21 (Staircase C, Level 4)

Task Description

We can use datasets on formality, politeness, offensive language, toxicity prompts, or others we find interesting to explore. We hope to show improved generation control while maintaining fluency with automatic and human evaluations. One potential foreseeable difficulty or subproblem is memory usage for stored encodings. This can be mitigated by only storing encodings of sentences with relevant attributes and could be an interesting direction for further exploration. Work can be submitted for publication upon completion.



References

[1] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations (ICLR)*, 2020.

[2] Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*, 2018.

[3] Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. Ruddit: Norms of offensiveness for English Reddit comments. In *ACL*, 2021.

[4] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *EMNLP*, 2020.

[5] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In *ACL*, August 2013.