FUDGE: Controlled Text Generation With Future Discriminators

Kevin Yang and Dan Klein
UC Berkeley

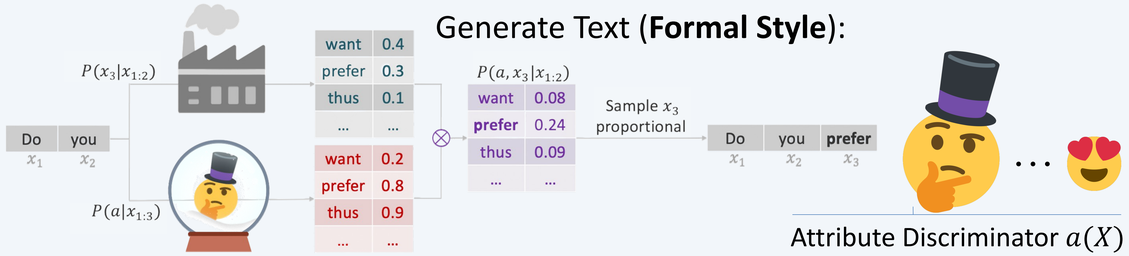
Controlled Text Generation

Language Model $P(X)$

Attribute Discriminator $a(X)$

For a desired attribute a (e.g. formal style), controlled text generation samples from the distribution $P(X|a(X) = True)$, or henceforth $P(X|a)$.

Guiding Generation Via Bayesian Factorization

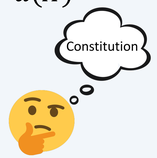


(1) Base language model predicts next token probabilities $P(x_{i+1}|x_{1:i})$ (turquoise table).
(2) Future-aware discriminator predicts $P(a|x_{1:i+1})$ for each possible next token (red table).
(3) Multiply (1) and (2), proportional to $P(x_{i+1}|x_{1:i}, a)$ by Bayes' Rule (purple table).
(4) Sample next token according to (3); repeat until generation complete.
Note: FUDGE can easily handle multiple intersected attributes by using multiple discriminators.

Future Discriminators

Generate Text (**Use "Constitution"**):

Non-Future-Aware Discriminator $a(X)$



Grep for "constitution"

Future-Aware Discriminator $P(a|x_{1:i+1})$

Does this continuation make "constitution" more likely later?

Evaluation Tasks

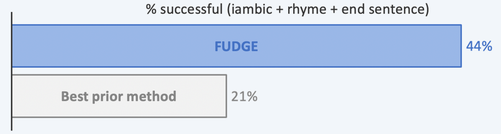
Uncontrolled Examples

FUDGE Examples

Results

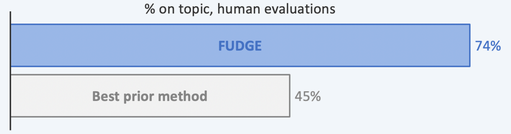
Complete a Poem

	And even thence thou wilt be stol'n, I fear,
Language Model	And even thence thou wilt be stol'n, I fear, and thou art a good friend of mine. The King's Guard…
FUDGE	And even thence thou wilt be stol'n, I fear, for this shall be the end. That's pretty clear.



Generate Topical Document (Politics):

	The issue focused on
Language Model	The issue focused on the role of a small company called Korn Ferry, which operates two vessels, the Korn and…
FUDGE	The issue focused on whether the two institutions were operating within the bounds set by the constitution…



Translate to English (Formal Style):

	[la información que que tenemos todo es propaganda entonces es portante ver otros versiones de lo que está pasando en el mundo]
Language Model	the information that, that we have is all propaganda, then is important to see other versions of what's happening…
FUDGE	The information that we have is all propaganda, so, it's important to see other versions of what's happening…