ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Creating a Topic Model of Language Models

Master's Thesis

Franz Knobel

knobelf@student.ethz.ch

Rycolab
Computational Linguistics, Natural Language Processing and Machine Learning
ETH Zürich

Supervisors:
Clara Meister
Prof. Dr. Ryan Cotterell

December 5, 2021

Acknowledgements

I thank Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Abstract

The abstract should be short, stating what you did and what the most important result is. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Contents

Chapter 1

Introduction

Todo: rewrite and complement

Pre-trained language models serve as the building blocks for many tasks in natural language processing (NLP). They have not only yielded state-of-the-art performance on myriad task, but have also reduced the computational resources often required to train NLP models by providing an advanced starting point for practitioners. Yet, due to the sheer size of these models, the learned probability distribution over natural language strings is difficult to analyze. The support of the distribution alone—the set of all possible strings that can be built using a specific vocabulary—is countably infinite. While a number of techniques have recently been proposed for analyzing the attributes of natural language that these models learn, it is still unclear what portions of the semantic space they learn (or fail) to represent.

Here we propose bringing back a standard model from natural language processing, the topic model, in order to gain a better understanding of this subject. By sampling strings from a pre-trained language model, we generate a pseudo-corpus that should provide an unbiased representation of the information learned by the model. We then learn topic models—using techniques such as LDA—to understand the distribution over topics that the model captures. Using this analysis technique, we hope to gain a better understanding of the variation in downstream performance across pre-trained models. We next propose a method for using these topic models to aid in an important downstream application of pretrained language models: natural language generation.

Recently, pre-trained language models have been applied to a number of text generation tasks, such as Abstractive Summarization or Story Generation. These tasks typically require fine-tuning the model on some task-related dataset—otherwise, sampling (unconditionally) from the language model would generate random text, likely irrelevant text. We propose trying to control text generation using the topic model learned for the pre-trained model: given a chosen topic, we use the distribution over words learned by the LDA model, interpolating it with the distribution from our language model, in order to steer generation towards that topic. Such a method would avoid the used of computational resources required to fine-tune such models, and will hopefully make controlled generation techniques produce more natural text.

In this thesis, the student will employ a combination of techniques from "*Topic Modeling*", "*Pre-trained Language Models*" and "*Topic-guided Text Generation*". The first part of this thesis will consist of learning and analyzing the different topic distribution that pre-trained language models capture. The second part of this thesis will involve analyzing these distributions and the effects they have on various model attributes and performance. Lastly, the resulting topic distributions will be used to try to guide text generation from pre-trained language models towards a specific topic.

Chapter 2

Related Work

Chapter 3

Problem Statement

Chapter 4

Analysis

Chapter 5

Solution

Chapter 6
Experimentation

preliminary study: train your own language model on a dataset, then run LDA on both that dataset and on the text generated from the language model and see how the two topic models compare

look into probing

Chapter 7

Evaluation

Chapter 8

Conclusion

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

Todo: This is a TODO annotation.

8.1 First Section Title

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

8.1.1 First Subsection Title

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

Theorem 8.1 (First Theorem)**.** *This is our first theorem.*

Proof. And this is the proof of the first theorem with a complicated formula and a reference to Theorem 8.1. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.

$$\frac{\mathrm{d}}{\mathrm{d}x}\arctan(\sin(x^2)) = -2 \cdot \frac{\cos(x^2)x}{-2+(\cos(x^2))^2} \tag{8.1}$$

□

And here we cite some external documents [1, 2]. An example of an included graphic can be found in Figure 8.1. Note that in LaTeX, “quotes” do not use the usual double quote characters.

Figure 8.1: This is an example graphic.

Bibliography

[1] A. One and A. Two, “A theoretical work on computer science,” in *30th Symposium on Comparative Irrelevance, Somewhere, Some Country*, Jun. 1999.

[2] A. One and A. Two, “A theoretical work on computer science,” in *30th Symposium on Comparative Irrelevance, Somewhere, Some Country*, Jun. 1999.

Appendix A

Code something