CSE 842 Project Proposal

Jeremy Arsenault

September 2022

1 Overview

What has motivated this project proposal is an interest in properties intrinsic to language and how they can be used for natural language tasks. Specifically, shannon entropy [1] will be the primary focus of this project, as it is a quantity which is well known to the fields of statistical linguistics and computer science. Also, it intuitively appears to be related to many interesting semantic qualities of language.

(PART 1 Q2-4) The scope of this project separates naturally into two distinct components. The first component deals with hypotheses related to language and information. In it I will consider a class of hypotheses, consider the different methods available to us for answering them, and use these to (hopefully) identify interesting relationships between language and entropy. **Example: In a question answering setting, entropy of a response correlates to its degree of difficulty (topic taught to first grader vs. college student).**

The second component deals with the application to natural language tasks of the insights gained from the hypotheses from the previous component. Of particular interest is application to text summarization and question answering tasks. If the relationship identified in the first component is sufficiently strong, it may be possible to leverage in search to control the generation of qualitatively diverse text using the same language model.

Key Point: Although entropy is a quantity which can be ascribed to a language as a whole, it is possible to use a language model to compute local entropy of a sequence, or the entropy of the set of sequences conditioned on some task-related variable. The latter is what I plan to investigate.

2 Prior Work

(PART 3 Q1+2) Idea of using entropy as a linguistic tool dates back to the origin of its modern formulation [1]. Similarly, for as long as people have been building statistical language models, perplexity and cross entropy have been used as a measure of model quality. This can be seen as related to a

specialization of entropy under strong simplifying assumptions. More recently, authors have proposed evaluating models using a more careful treatment of entropy [2]. Still though, entropy is generally considered as a property intrinsic to an entire language [3], and of an entire language model respectively.

More explicitly - I'm aware of lots of work very closely related to the problem of this project including: Statistical approaches [1] to calculate the entropy of a language. Evaluating language models (statistical and ML) using perplexity and cross entropy [2]. Evaluating ML language models using entropy-based metrics [2].

(PART 3 Q3) Clear limitations to classical statistical computations of entropy (as discussed in class). Also there criticisms of perplexity-per-word metric. As I don't plan on building an entropy based evaluation metric these are not immediately relevant. See [2] for more details on this problem

(PART 3 Q4) What exactly does this mean? As I see it, entropy of a language across a diverse population isn't a well defined quantity. Communication happens between individuals! The way these individuals use language depends on their prior knowledge, the assumed prior knowledge of the target audience, and many other factors like dialect and age[4]. To be clear, entropy can certainly be defined and measured at the level of an entire language, but the quantity may also be subject to interesting intra-language variations.

(More explicitly - results of these are qualitatively interesting, but difficult make statements like 'this entropy calculation is better' or 'this evaluation metric is better'. Further, the work cited above does not address the key question of this project as it treats entropy as a quantity intrinsic to a language as a whole.)

In this project, the focus will be on identifying these variations and drawing insights from them. To my knowledge, there is a very limited amount of literature on this topic (if you know of any that you can share with me, it would help a ton!

3 Hypothesis Testing

In light of this, many questions naturally arise related to the semantic nature of these variations in entropy. For instance: What is the relationship between a (communicator / audience member)'s (age / expertise / subject / task) and the entropy of that communication?

(PART 3 Q5) Answering these questions is not a trivial endeavor, but essentially hinges on access to data and a computationally tractable approximation of entropy. We need a way to measure the entropy of these categories, but obviously it isn't practical (or possible) to create a language model for each class for each hypothesis. Given a sufficiently expressive foundation language model, I propose that by conditioning the generation on text that is characteristic of a given class, we can approximate that classes respective entropy. Language models implicitly model the mannerisms of individual communica-

tors, and as such essentially 'contain as submodels' everything we need to test a hypothesis.

(More explicitly - I plan on using a pre-trained MLM foundation model like BERT to make conditional entropy calculations.)

Still need to try to formalise and develop these ideas a bit more :)

4 Data

(PART 2 Q1:) The data requited for this project is going to be hypothesis specific. For instance, in the example hypothesis in the overview section, I have scraped transcriptions from lectures on youtube and coursera.

5 Applications to Natural Language Tasks

A) Use entropy-related quantity to characterize text (ex. difficulty of material). (This will be the main one!)

B) Naive generation. Use (A) combined with any language model for generation. Sample sequences, choose sequence with desired properties.

C) Principled generation. Impose entropy-related constraint / objective directly on search to achieve B).

(PART 3 Q5:) I plan on directly using a pre-trained question answering or text summarization model. Sequences can be generated naively then characterized after. Alternatively, a modification of the sequence search algorithm (beam, bfs) which operates under some entropy constraint / objective can be proposed.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x

[2] K. Arora, "Contrastive perplexity: A new evaluation metric for sentence level language models," *CoRR*, vol. abs/1601.00248, 2016. [Online]. Available: http://arxiv.org/abs/1601.00248

[3] C. Bentz and D. Alikaniotis, "The word entropy of natural languages," 2016. [Online]. Available: https://arxiv.org/abs/1606.06996

[4] M. Brysbaert, M. Stevens, P. Mandera, and E. Keuleers, "How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age," *Frontiers in Psychology*, vol. 7, 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01116