Limitations of entropy in n-gram language modelling

Even though modern language models are usually trained in neural network infrastructures, like a recurrent neural network (Mikolov, 2012) or a transformer (Radford et al., 2019), the more classical probabilistic n-gram language model is still valuable because of its simple and straightforward nature. As this type of model is easy to interpret, it is often used to explain basic concepts of language modelling (Jurafsky & Martin, 2021). Additionally, this type of language model is a very good option for a data-driven, descriptive and exploratory analysis of a dataset as its bottom-up approach can easily identify sequential patterns typical for the data.

Both for the intrinsic evaluation of these n-gram models and the probability estimation of any given text input, entropy (and its derived perplexity) is commonly used. This measure, based on Shannon (1948), gauges how uncertain an input sequence is according to the probability distribution of the n-gram language model. A lower entropy (regularly expressed in bits) suggests that the sequence is less uncertain because less information is needed for its prediction. This would therefore imply that text sequences with lower entropy are more probable and therefore more typical for a language model – as they are easier to predict.

A practical application of this measure is to indicate how typical a subset of data is with regard to the entire dataset. For example, one could measure how probable a paragraph is, according to the language model of an entire book. The more probable the paragraph, the more representative it is for the rest of the book. When using the original Shannon entropy, however, this does not always hold true and can produce deceiving, yet correct, probability estimations. Two main properties of this estimation play a role in this unexpected output: (i) the logical influence of text sequence length on probability estimation, and (ii) a possibly unwanted impact of disproportionally high probabilities in language models with lower n-gram counts. While this first property is not unknown, the second one is often overlooked, as it is an issue which occurs mainly in smaller datasets. Instead of applying the Shannon entropy measure 'as is', an adapted version is required to compensate for these limitations and achieve the previously mentioned goal of identifying text sequences typical for the model.

Within this contribution, the aim is to closely examine entropy and present a clear and practical overview of the assumptions of entropy, its possible limitations, and ways in which the measure can be adapted to compensate for these possible limitations.

Jurafsky, D., & Martin, J. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

Mikolov, T. (2012). *Statistical Language Models Based on Neural Networks* [PhD Thesis, Brno University of Technology].

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. 24.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.