Introduction to Neural and Cognitive Modelling

Project Proposal

Theme: Bayesian Language Modelling for Language Generation

For our project, we plan to obtain a deeper understanding of how language is acquired from the surroundings and how a fundamental computational unit models human language. After going through the series of provided lectures related to language and cognition, we now have a decent understanding of the key processes and the theoretical intuition behind how the human brain facilitates human language and how it can be modelled using fundamental statistical techniques.

We would aim to broadly achieve the following two goals out of our Project

A Formal Study

It is clear that there are two schools of thought: Chomskyan and Greenbergian - when it comes to defining how the human mind acquires natural language from its environment. These two schools of thought have their associated arguments, many of which we have seen in this given course. We aim to conduct a formal study of these two schools of thought, and present a theoretical analysis and comparison of the two ideals, tracing the story from the understanding of language itself to its modelling and its process of acquisition.

An Implemented System

The main aspect of our project will be Bayesian Language Modelling and its application in the generation of natural language sentences. Chomsky's generative grammar theory allows one to present a CFG (Context Free Grammar) from which all natural language expressions can be derived. Greenberg however, indicates that language is acquired from one's surroundings by perceiving the natural language expressions and dialogues, with the background of specific contexts and social scenarios. In this, an understanding of which word or phrase to use in which setting plays a very important role.

Chomsky's poverty of stimulus argument hints at the existence of a grammar and the innateness of human-beings to have a faculty which allows them to acquire language. Modern medical science research has clearly shown how Broca's and Wernicke's areas of the human brain are instrumental for human language, which is something we saw in our course. But, it is clear that an understanding of semantics is required and the knowledge of syntax itself is not enough for the generation of 'meaningful' sentences. In this regard, consider the well-known example of "Colourless green ideas sleep furiously", which though syntactically correct, makes no sense at all. Coming to the present day, when we look at language on the internet on platforms such as Twitter, formally defined CFG rules as outlined in old linguistics textbooks and the lecture slides of Chris Lucas do not work out for validating sentences or creating new content. Therefore, the modelling approach involving a well defined CFG as an intermediate is not really scalable to the ever-growing and increasingly adaptive domain of natural language in practical day-to-day usage. In this regard, we fully fixate on the approach of Bayesian Language Modelling.

We aim to implement an *N*-gram language model with modern devices for error handling. We also aim to incorporate relatively new schemes such as interpolation, backoff and smoothing to ensure robustness of our language models. We'll train our learning model on large corpora to obtain probabilities of *N*-grams. In terms of a goal based learning objective, we aim to minimize the model's perplexity scores and maximize the model's entropy levels. This is a purely unsupervised task which ensures that the model amplifies the occurrence probabilities of the observed sentences and phrasal units. An expectation-maximization schema captures the iterative parameter update. At the end, this language model can be leveraged to create a sentence generator module.

If time persists, we also aim to create a secondary language model for Twitter data and a Tweet generator *purely using Bayesian Language Modelling*. We'll evaluate our approaches mostly using Perplexity scores as defined in the provided course material. In terms of implementation details, for a clearer understanding of how Bayesian Modelling works from the very first principles, we'll refrain from using any pre-existing libraries and implement modelling objectives using basic Python3 only.

Primary Reference Papers

1. Bayesian Estimation Methods For N-gram Language Model Adaptation
2. Rules and Similarity in Concept Learning
3. Bayesian Language Model based on Mixture of Segmental Contexts for Spontaneous Utterances with Unexpected Words
4. A Hierarchical Bayesian Language Model based on Pitman-Yor Processes
5. Scalable Bayesian Learning of Recurrent Neural Networks for Language Modeling
6. A primer on probabilistic inference

Further Resources

1. Reading list on Bayesian modeling for Language
2. Lectures on Bayesian Modelling for Natural Language by Chris Lucas at the University of Edinburgh
3. Speech and Language Processing Textbook