**Proposal -**

Hierarchical-XLNet: Exploiting compositionality for permutation language modeling

Xuan Chen

Student # 82286956

Tianyu Hua

Student # 38066445

# Introduction and Motivation

This project introduces Hierarchical-XLNet (HXLNet), a new language model that incorporates a hierarchical structure into the permutation language modeling approach. HXLNet builds on the advancements made by BERT and XLNet to address their limitations and improve the performance of language models across various NLP tasks. In particular, HXLNet aims to better capture the structure of the input sequence by predicting the probability distribution across all possible permutations of the hierarchical structure.

In this proposal, we will provide a detailed overview of the language model based on permutation, including its strengths and weaknesses. We will also describe the research design of existing permutation language models and highlight their contributions to the field. Additionally, we will present our expected contributions with HXLNet and discuss how this model can overcome the limitations of current permutation-based language models. Finally, we will provide details on the dataset we will use for evaluation and specify the evaluation metrics we will employ to assess the performance of HXLNet.

# Related Literature

In recent years, pre-training and fine-tuning NLP processing solutions have become increasingly popular, and two common pre-training methods are autoregressive (AR) and autoencoding (AE) models. AR models estimate the probability distribution of a text corpus based on the conditional probability of text sequences from left-to-right and right-to-left, but they are limited in their ability to model deep bidirectional contexts.

To address these limitations, permutation language modeling was introduced in the paper "XLNet: Generalized Autoregressive Pretraining for Language Understanding", which enables the model to capture dependencies between all positions in the input sequence without bias toward a specific direction. XLNet builds on state-of-the-art language models such as BERT and GPT-2, which suffer from the bias problem.

XLNet's permutation language modeling predicts the probability of a word at a specific position in the sequence given all other words in the sequence, regardless of their order (Figure 1). This allows the model to learn from all possible orders of the input sequence and capture dependencies between all positions in the sequence. XLNet is a generic autoregressive method that takes full advantage of the benefits of AR and AE, avoiding their limitations.

In addition to proposing a new pre-training goal, XLNet also improves the design of the pre-training framework by applying segment recurrence in Transformer-XL and the relative encoding scheme to pretraining, which is particularly significant on long text sequences. The Transformer(-XL) network is also re-parametrized to reduce uncertainty in permutation-based language modeling.

Compared to BERT, XLNet is able to capture more dependencies between tokens and contains more training information. A standard AR language model like GPT is only able to cover partial dependencies, while approaches like ELMo lack deep interaction modeling in both directions.

In this proposal, we will provide a detailed overview of permutation-based language models, including their strengths and weaknesses, describe the research design of existing permutation language models, and highlight their contributions to the field. We will present our expected contributions with HXLNet and discuss how this model can overcome the limitations of current permutation-based language models. Finally, we will provide details on the dataset we will use for evaluation and specify the evaluation metrics we will employ to assess the performance of HXLNet.

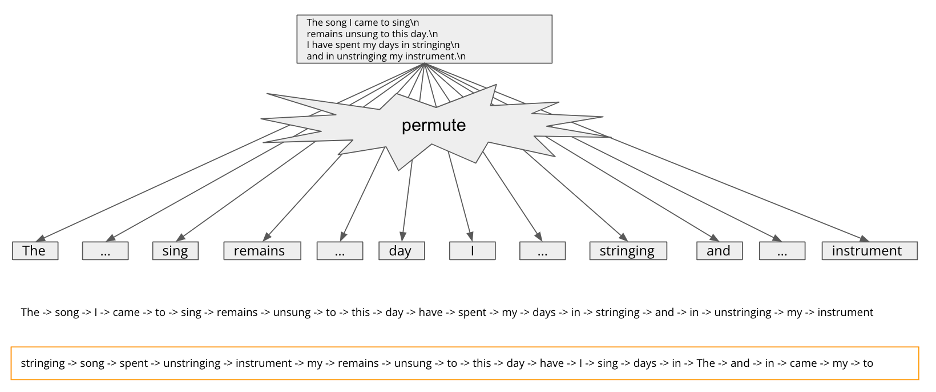


Figure 1: Permutation Language Modeling

# Introduction to Hierarchical Language Modeling

In natural language, hierarchies refer to the organization of concepts or ideas into a structure of levels or tiers based on their relative importance or relationships. One common type of hierarchy is a tree structure, where a concept at the top serves as the root node and branches out into sub-concepts or sub-topics that are more specific or detailed. For example (Figure 2), in the graph “The song I came to sing remains unsung to this day. I have spent my days in stringing and in unstringing my instrument.”, we regard the root as:

The song I came to sing\n

remains unsung to this day.\n

I have spent my days in stringing\n

and in unstringing my instrument.\n

which then branches out into sub-nodes as two separate sentences.

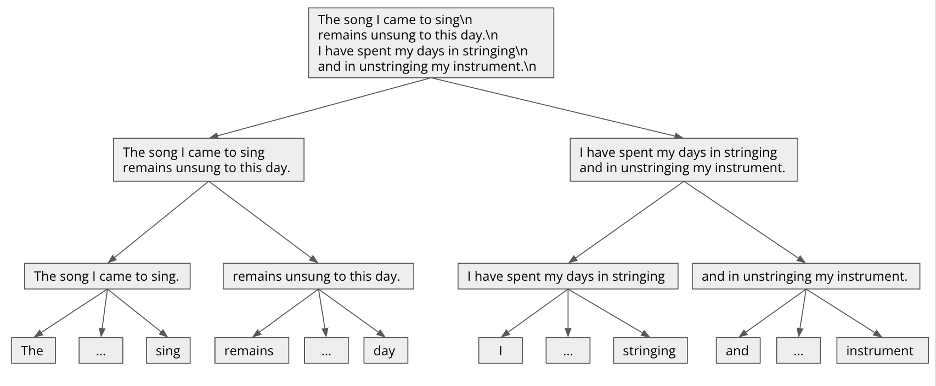


Figure 2: Hierarchies in Natural Language

To improve the current permutation language modeling, we consider integrating hierarchical language modeling. We are going to introduce a hierarchical structure to the input sequence and train the model to predict the probability distribution over all possible permutations of the hierarchical structure. Hierarchical language modeling is a type of language modeling that incorporates a hierarchical structure into the input sequence. The idea behind hierarchical language modeling is to break down the input sequence into smaller, more manageable parts (Figure 3), which can be modeled independently and then combined to form a complete representation of the sequence. It will capture dependencies at different scales of natural language.

In this project, we would propose a permuting approach that respects the hierarchical structure of input language sequences. And also, we would demonstrate the superiority of our pretraining method compared with permutation language modeling in finetuning on downstream commonsense tasks.

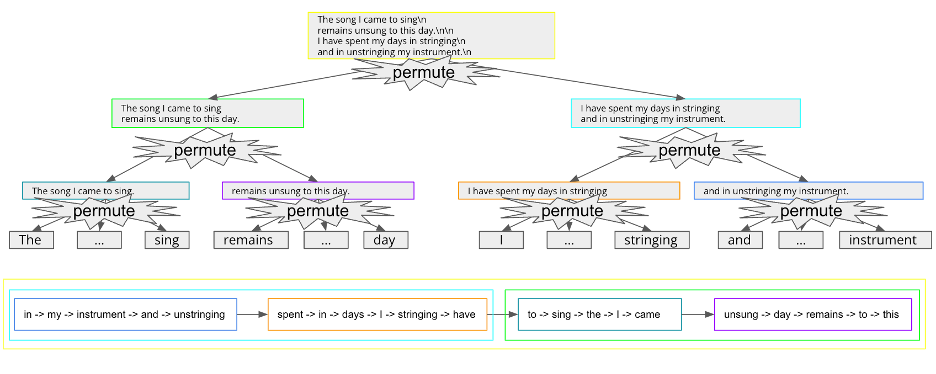


Figure 3: Hierarchical Language Modeling

# Dataset

We are going to use **Wikipedia-en**, which is an English language version of Wikipedia, the world's largest online encyclopedia, with over 6 million articles and 3.7 billion words.

- Pretraining: HXLNet is pretrained on text sequences using hierarchical language modeling, we input text hierarchically permuted and model their joint distribution of a given sentence.

- Finetuning: The pretrained HXLNet can be fine-tuned to downstream tasks. During fine-tuning, additional layers can be added to the Transformer with randomized parameters: these parameters and those pretrained Transformer parameters will be updated to fit training data of downstream tasks. We will use the COPA dataset and follow the same experiment setting in assignment2. We will compare the performance of the pretrained model of permutation language modeling and hierarchical language modeling.

# Evaluation Metrics

To evaluate the performance of the fine-tuned HXLNet model on downstream tasks, we conduct experiments on the COPA dataset using the same experiment setting as in assignment2. We compare the performance of the pretrained model of permutation language modeling and hierarchical language modeling.

We use the following procedures to measure the model's performance on the COPA dataset:

1. Compute the validation predictions. This is not in the original notebook, but can be done with the trainer.predict function.
2. Sample 25 examples that the model predicted incorrectly (if your model predicted less than 25 examples incorrectly, analyze all the errors).
3. Manually categorize the error types and report the percent of each error along with an example.

# Division of tasks between group members and timeline

1. Timeline

* 14/2 - 15/2: Project discussion
* 20/2 – 1/3: Project proposal presentation and written proposal
* 6/3 – 13/3: Experiment and adjust model
* 15/3 – 27/3: Final project presentation preparation
* 29/3 – 8/4: Final project report

1. Division of tasks of members

* Tianyu Hua: preprocess dataset, fine-tuning, structure model architecture
* Xuan Chen: research the original paper, analyze model architecture, and process dataset.

# 