NLP Group Assessment:
Understanding and Generating Explanations from the RuozhiBa Dataset

JC4003 NLP

1 Assessment Overview

In this group assessment, you will explore and experiment with traditional machine learning and deep learning models, including large language models (LLMs), to generate accurate meanings and explanations for the samples provided in the RuozhiBa dataset. The purpose of this exercise is to apply your knowledge from the course to a real-world dataset, practicing your skills in data annotation, model design, and evaluation.

2 Objectives

2.1 Data Annotation

Each student will be responsible for annotating a portion of the RuozhiBa dataset **in Chinese** to provide **clear explanations** for each sample. This will aid in understanding the actual meaning behind the data samples.

2.2 Model Training

Working in groups, you will build models to **generate accurate meanings or explanations for unseen data samples** using the annotated dataset. You may choose from traditional machine learning methods, deep learning models, or large language models.

2.3 Model Evaluation

Your model's performance will be assessed using both **automatic evaluation metrics**, such as BLEU and ROUGE, and **human evaluation** conducted by your group.

2.4 Presentation & Report

Each group will present their model, explain their design choices, and demonstrate their model' s performance. Additionally, you will submit a detailed report on your process.

3 Steps & Requirements

3.1 Data Annotation

The dataset will be divided, and each student will receive a subset for annotation with everyone's student ID as filename. **You should annotate each data sample in Chinese, explaining its meaning in clear and concise terms.** These annotated samples will be combined to form the final dataset, which will be split into training and test datasets.

Here are some examples to help you understand better:

Example 1:

- Original data: 根据牛顿第一定律，我推算出本次世界百大物理学家排名，爱因斯坦只能屈居第二
- Annotated result: 牛顿第一定律本意指的是牛顿提出的第一条被公认的物理定律，而不是牛顿排名第一的定律

Example 2:

- Original data: 浴霸打了一个响指，给全世界一半的人洗了澡
- Annotated result: 这里浴霸打响指借用了复仇者联盟中灭霸一个响指可以消灭全世界一半人口的概念进行类比，因此有了给全世界一半的人洗澡的结果

3.2 Forming Groups

You will form groups of 4-6 students. Each group will work collaboratively on building a model to generate the meanings for the data samples. Group formation is flexible within each programme (cross-programme group is not allowed), but must be completed by **Monday, September 23, 2024**. Each group leader should send the member list to the corresponding course coordinator after the deadline.

3.3 Model Development

You are free to use traditional machine learning models (e.g., Naive Bayes, Logistic Regression) or deep learning models (e.g., RNN, LSTM, Transformers). For those interested in LLMs, the recommended approach is to use *prompt engineering* techniques to guide the LLM in generating accurate meanings for the dataset samples. Groups that wish to challenge themselves can attempt to *fine-tune LLMs* using the annotated RuozhiBa dataset to improve their model's performance.

3.4 Model Evaluation

Each group's model will be evaluated using:

- **Automatic evaluation metrics**: such as BLEU, ROUGE, and other applicable metrics to assess the generated meanings' accuracy.
- **Human evaluation**: where your group will assess the quality of the outputs based on specific criteria (e.g., fluency, accuracy, relevance).

3.5 Presentation & Report

Each group will prepare a **presentation** to explain the design of their model, demonstrate its performance on the test dataset, and discuss challenges faced and solutions implemented.

You will also submit a **detailed report** that covers:

- **Introduction & Objectives**: Why you chose your model(s) and what you aimed to achieve.
- **Methodology**: A step-by-step explanation of your approach, from annotation to model design and training.
- **Experiments & Results**: Your evaluation results, observations, and any adjustments you made to improve your model.
- **Discussion**: Insights, challenges, and future work you would consider.

The report should be between **3000-5000 words** and include references to the tools, libraries, and models you used. **Each group member's contribution and percentage should be highlighted at the beginning of the report.**

4 Evaluation Criteria

4.1 Data Annotation (20%)

Goal: Evaluate the clarity, accuracy, and comprehensiveness of the annotated explanations for the RuozhiBa dataset samples.

Criteria	Weight	Description
Clarity of Explanation	5%	The annotations should provide clear, easily understandable explanations of each sample's meaning. No ambiguity or vagueness should be present.
Accuracy of Annotation	5%	The meaning of each sample should be annotated correctly in line with its context. This involves capturing nuances and key elements accurately.

Consistency of Terminology	5%	The use of terminology should be consistent throughout the annotations, especially when describing similar concepts across different samples.
Completeness	5%	All samples in the assigned portion of the dataset should be annotated. No gaps or skipped samples should be present.

4.2 Presentation (40%)

Goal: Assess how effectively the group explains their methodology, model design, and results in a clear, professional manner.

Criteria	**Weight**	**Description**
Introduction & Objectives	8%	The group provides a clear introduction to their approach, objectives, and rationale for choosing their models and methods.
Methodology Explanation	12%	Clear and logical explanation of the methodology. This includes the model design, choice of algorithms, training processes, and evaluation setup.
Results & Demonstration	12%	The group demonstrates their model' s performance on the test dataset. Includes a clear discussion of automatic and human evaluation metrics.
Visual Aids & Communication	4%	The presentation is well-organized, with clear slides, diagrams, or visual aids. The group communicates confidently and explains key points clearly.
Q&A Handling	4%	The group effectively handles questions from the audience or instructor, demonstrating understanding of their model and results.

4.3 Report (40%)

Goal: Assess the depth of the group' s understanding, analytical rigor, and ability to communicate their work in a structured, professional format.

Criteria	**Weight**	**Description**
Introduction & Objectives	5%	The introduction clearly defines the group's goals, the problem they are solving, and the approach they are taking.

Dataset & Annotation Process	5%	Clear explanation of the RuozhiBa dataset and the group' s approach to annotation. Discussion of challenges faced during annotation (if any).
Model Selection & Methodology	10%	Detailed description of the model(s) selected, the rationale for choosing them, and the methodology used. Includes data preprocessing steps, model architecture, and training processes.
Experiments & Results	10%	Presentation of experiments conducted, results obtained (using BLEU, ROUGE, and human evaluation). Includes insightful analysis of the results, including error analysis and discussion of challenges.
Discussion & Critical Analysis	5%	Discussion of the model' s strengths, weaknesses, and potential improvements. Includes a critical evaluation of why certain decisions were made, including potential trade-offs.
Future Work & Improvements	3%	The group provides a thoughtful discussion of how the work could be extended or improved in the future.
Structure, Writing & Clarity	2%	The report is well-structured, with clear sections, proper use of headings, and professional writing. No major grammatical or spelling errors.

4.4 Final Breakdown

- **Data Annotation:** 20%
- **Presentation:** 40%
- **Report:** 40%

5 Deadlines

- **Group formation**: Monday, September 23, 2024
- **Data annotation submission**: Tuesday, September 24, 2024
- **Presentation day**: AI: Friday, October 18, 2024; CS & BMIS: Monday, October 21, 2024
- **Report & Code submission**: 22:00 Tuesday, November 5, 2024 (SCNU Time)