Relatório do Trabalho Final

Disciplina: Linguıstica Computacional

Aluno: Ariel Augusto dos Santos

Introdução e Motivação do Problema

O objetivo deste trabalho foi a criação de um modelo para geração de paráfrases de sentenças em português. A motivação para esta aplicação na realidade era a intenção original do trabalho: reproduzir o método **STRAP** (Style Transfer via Paraphrasing) para transferência de estilo não supervisionada, proposta em 2020 por Krishna et al, na língua portuguesa (ao invés de inglês como no *paper* original). Porém, como o nome indica, este método depende do treinamento de um modelo para paráfrases, e não encontrei nenhum trabalho ou grande datase voltado a esta tarefa para o português. Daí decidi criar um parafraseador baseado em um modelo de linguagem multilíngue.

Dados

O maior *dataset* de paráfrases que encontrei em português foi o TaPaCo Corpus:

> A freely available paraphrase corpus for 73 languages extracted from the Tatoeba database.

Ele possui 78430 sentenças em português, agrupadas em 29949 grupos aproximadamente equivalentes.

Na preparação do *dataset*, todas sentenças são agrupadas por equivalência e então separadas em tuplas, gerando 221306 tuplas de duas sentenças (filtrando-se tuplas com ambas sentenças idênticas).

Modelo

Modelo de Linguagem Pré-Treinado

Utilizei o modelo T5: *Text-To-Text Transfer Transformer*, proposto em 2020 pelo Google (Raffel *et al*, 2020 - *Journal of Machine Learning Research*), especificamente a versão multilíngue publicada em 2021 por Xue *et al*, o **mT5**.

O checkpoint utilizado é o `mT5-Base`, com 580 milhões de parâmetros. Com o dataset TaPaCo preparado, executamos uma época de treinamento para *fine-tuning* com a taxa de aprendizado de 1×10^{-4} e o *optimizer* AdamW.

Por questão de recursos de memória, limitamos as entradas e saídas ao tamanho de 72 tokens e processo os dados em batches de 10.

Resultados

10% do *dataset* foi separado para avaliação do modelo, e após uma época de treinamento aparenta convergir em perdas de $\approx 2, (\pm 0.1)$ (em termos de *Binary Cross Entropy*).

Exemplos de Saída

Entrada	Saída
A vontade geral deve emanar de todos para ser aplicada a todos.	A vontade geral deve emanar de todos.
Se queres prever o futuro, estuda o passado.	Se queres prever o passado.
Todos os seres humanos nascem livres e iguais em dignidade e em direitos.	O seres humanos nascem livres.

Avaliação Subjetiva e Conclusão

O modelo não é capaz de preservar a semântica original das frases, e geralmente se limita a encurtar as sentenças (abandonando as palavras no meio). Especulo que isso é causado em grande parte pela limitação do dataset de treinamento: enquanto o paper de Krishna et al utiliza o ParaNMT-50M (Wieting & Gimpel, 2018), com 50 milhões de **pares** de paráfrases, só tive 78430 sentenças à minha disposição e muitos pares consistiam de variações triviais.

Possíveis trabalhos futuros

- experimentar com outros *optimizers*, como SGD
- otimização do T5 para minimizar custo de memória e viabilizar o uso de versões maiores que o `mt5-small`, filtrando vocabulário de tokens, ajustando hiperparâmetros, etc
- aquisição/extração/geração de *datasets* maiores de paráfrases.
- experimentar com diferentes modelos de linguagem treinados para português, por ex. BERTimbau