SemEval 2024 Task 1: Semantic Textual Relatedness

Thad Greiner
CSE Department
Michigan State University
greine30

Yufeng Li
CSE Department
Michigan State University
liyufen1

Abstract

In this project, we will focus on the first task of SemEval 2024: Semantic Textual Relatedness (STR). Our goal is to build robust STR models using a multilingual dataset that includes nine different languages with corresponding tags. We will apply a combination of traditional machine learning algorithms, such as neural networks (NN), Random Forests, Support Vector Machines (SVMs), and Logistic Regression, alongside deep learning approaches like mBERT, LaBSE, and XLM-RoBERTa for feature extraction and preprocessing. For evaluation, we will utilize a provided test script, as well as our own metrics, including Mean Squared Error and Pearson correlation, to assess the performance of our models.

1 Introduction

Semantic Textual Relatedness (STR) is an important task in natural language processing (NLP) that measures the relatedness between two different texts across multiple languages. This project focuses on the first task from SemEval 2024, which aims to develop robust STR models using a multilingual dataset comprising nine distinct languages. The significance of this research is underscored by the importance of semantic understanding in a multilingual context. Currently, much of the related work concentrates on widely spoken languages, such as English. However, STR can be applied to understanding and learning various languages, including those at risk of disappearing. This capability will also benefit other fields, such as history, art, social sciences, and language preservation, by enhancing our understanding of different languages and cultures.

2 Related Works

Semantic Textual Relatedness (STR) has a long history in NLP research. Early work explored the concept of semantic cohesion, with Halliday and Hasan (1976) introducing the idea of cohesion in English text. Their work provided new methods for linguistic analysis and laid the foundation for future studies in semantic similarity.

More recent research has focused on the creation of datasets to advance STR studies. Asaadi, Mohammad, and Kiritchenko (2019) introduced Big BiRD, a large, fine-grained bigram relatedness dataset that serves as a valuable resource for examining how semantic composition influences STR.

In 2023, Abdalla, Vishnubhotla, and Mohammad introduced the STR2022 dataset, which further explores sentence-level semantic relationships. Their findings demonstrate the dataset's efficiency in various NLP tasks, highlighting its utility for evaluating semantic relatedness in diverse contexts.

3 Methodology

The goal of this project is to predict semantic textual relatedness between sentence pairs in a variety of languages via supervised learning, with each dataset provided by the SemEval group. As such, we will have several stages to go through: data preprocessing, feature extraction, model training, and evaluation. Given the multilingual nature of this project, we will need to ensure our techniques are applicable for each language. Below are our current plans for each part of the process:

3.1 Data Preprocessing

As the dataset for each language consists of two sentences and a similarity score, we will need to convert the data to a more usable form. For this, we will use tokenization, normalization, and padding before feature extraction. While this process will be relatively straightforward for English, we will need to account for language-specific differences during tokenization and normalization.

3.2 Feature Extraction

After basic preprocessing, we will utilize multilingual models such as mBERT, XLM-RoBERTa, or LaBSE to transform each sentence into its vector representation. These models have been trained on large corpora across multiple languages, making them suitable for our purposes. From there, we plan to explore various methods for combining sentence embeddings to effectively highlight similarities, including concatenation, absolute difference, averaging, cosine similarity, and pooling. These should help capture diverse semantic relations across the languages we are training on.

3.3 Model Training

Once we have obtained our features as desired, we will need to train a model to predict a continuous value that reflects sentence similarity. To do so, we will initially start with a simple regression model like logistic regression before progressing to more advanced models such as random forest regression, neural networks, and SVMs. Since there are multiple languages, we currently plan to train a separate model for each language, though we may attempt to train a single model for all languages as an additional goal. Furthermore, k-fold cross-validation will be employed to ensure the models perform robustly.

3.4 Evaluation

For evaluation, we will use the testing data and scripts provided to assess predictions, which mainly use the Spearman rank correlation. Additionally, we will implement our own evaluation metrics, including Mean Squared Error and Pearson correlation, to assess how our predicted similarity scores compare to the true values.

3.5 Additional Goals

Provided we have time, there are a couple of additional goals we would like to reach. These include more advanced hyperparameter tuning, training a model with all of the languages rather than separately, testing other models like GNN, investigating whether one language's model can aid another, and attempting the other two tasks from the SemEval challenge (unsupervised and multilingual tasks).

3.6 Summary

Overall, here is the methodology we propose for achieving the goal of predicting semantic similarity of sentences for a variety of languages. Going forward, we will implement these ideas and determine which techniques and models will perform the best, and hopefully achieve better results than those already recorded by others who have worked on this task.

4 Division of Work

- Thad Greiner:
  - Proposal: Worked on drafting the proposal and methodology.
  - Preprocessing Tasks: Implement tokenization and normalization for English and one additional language.
  - Feature Extraction Methods: Test different feature extraction methods (e.g., concatenation, pooling).
  - Model Training: Train models: SVM and Random Forest.
  - Evaluation and Comparison: Analyze evaluation metrics for the models trained.
  - Reports: Work together to create midterm and final reports.
- Yufeng Li:
  - Creating Draft: Collaborate on drafting the proposal and methodology.
  - Preprocessing Tasks: Implement tokenization and normalization for remaining languages.
  - Feature Extraction Methods: Test different feature extraction methods (e.g., cosine similarity, absolute difference).
  - Model Training: Train models: Logistic Regression and Neural Networks.
  - Evaluation and Comparison: Analyze evaluation metrics for the models trained.
  - Reports: Work together to create midterm and final reports.

References

[1] M. A. K. Halliday and R. Hasan. 1976. Cohesion in English. Longman, London.

[2] A. Asaadi, S. Mohammad, and V. Kiritchenko. 2019. Big BiRD: A large, fine-grained bigram relatedness dataset. In Proceedings of the 2019 Con-

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3207–3211.

[3] A. Abdalla, A. Vishnubhotla, and S. Mohammad. 2023. STR2022: A semantic textual relatedness dataset for various NLP tasks. In Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 200–210.

[4] Semantic Textual Relatedness (STR) Challenge. 2024. https://semantic-textual-relatedness.github.io/