Interactive Natural Language Inference Classification

Shira Rozenthal
Tulane University / New York, NY
shira.rozenthal@gmail.com

Peter Sapountzis
Tulane University / San Francisco , CA
peter.sapountzis@gmail.com

Problem Description

We'll be diving into the world of Natural Language Inference (NLI), which involves determining if a hypothesis is logically supported by a given premise. This task is pivotal for understanding textual relationships and categorizing them into entailment, contradiction, or neutrality.

For example, the premise "Peter and Shira are at the Tulane library working on this project"

Entails that Peter and Shira are on campus
Contradicts that Peter and Shira are in class
Is ***neutral*** on whether they are having fun

...(we're having fun we promise!) Our goal is to be able to take a user-inputted premise and hypothesis and classify the relationship between the two in real time.

Data

We are working with one of the leading datasets in this space, the MultiNLI (MNLI) dataset.

MNLI extends the concept of the SNLI dataset by incorporating a wider variety of text, including both spoken and written forms across ten different genres sourced from the Open American National Corpus.

These genres include transcriptions of face-to-face conversations, government documents, letters related to philanthropic fundraising, the 9/11 Commission report, non-fiction works published by Oxford University Press, articles from Slate Magazine, telephone conversations, travel guides, linguistics posts, and contemporary fiction. For each premise drawn from these sources, crowdworkers generated three new sentences to represent entailment (statements that are necessarily true if the premise is true), contradiction (statements that are necessarily false), and neutral (where neither condition applies). A distinctive aspect of MNLI is that only half of the genres are included in the training set, with the remaining unseen genres serving to test a model's ability to generalize to new text sources.

Method

For the project, we are utilizing the mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 model. This model is an advanced multilingual model capable of performing natural language inference (NLI) across 100 languages, making it highly suitable for multilingual zero-shot classification tasks. Developed by Microsoft, it is based on the mDeBERTa-v3-base architecture and has been pre-trained on the CC100 multilingual dataset, which encompasses text in 100 languages.

To evaluate the results of our implementation, we plan to use standard NLI and classification metrics such as accuracy, precision, recall, and F1 score, tailored to the specific tasks at hand. These evaluations will help us determine the model's performance in understanding and classifying multilingual text based on the context provided.

As baselines for comparison, we will look into the performance metrics of other multilingual NLI models that were state-of-the-art prior to the introduction of mDeBERTa-v3-base, such as XLM-R and multilingual BERT (mBERT). Comparing our results against these models will allow us to assess the improvements and advantages offered by mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 in handling multilingual NLI tasks.

Preliminary Experiments & Results

We have the dataset and mDeBERTa model loaded in a python notebook and have set up a naive classifier that's trained on the first 1,000 premise/hypothesis pairings. Our preliminary naive classifier has a 80.56% validation accuracy across the entire validation set.

Division of Labor

We've been working closely together in the early research stages. Moving forward, Shira will take the lead on fine-tuning the DeBerta model while Peter explores alternative approaches and builds the interactive interface for our final submission.

Related Works

A large annotated corpus for learning natural language

- This paper introduces the SNLI corpus, providing a foundational dataset for NLI research.

A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

- This paper introduces the MNLI corpus, the second largest NLI dataset available for use.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- This paper presents a method for enhancing deep contextualized word representations, boosting NLP task performance, especially NLI. It outlines a pre-train and fine-tune approach that improves our NLI model's sentence pair analysis.

Adversarial NLI: A New Benchmark for Natural Language Understanding

- This paper introduces the ANLI dataset, challenging NLU systems with adversarial examples to test model robustness. It's key for evaluating and improving NLI models against complex challenges, emphasizing adversarial testing's importance in developing stronger NLU systems.

IMPLI : Investigating NLI Models' Performance on Figurative LanguageLinks to an external site.

- The paper introduces the IMPLI dataset to evaluate NLI models on figurative language, revealing a significant gap in their ability to interpret idioms and metaphors. It shows advanced models like RoBERTa struggle in this area, highlighting the need for improved methodologies.

Timeline

We currently have a naive predictor running on 1,000 / 433,000 premise/hypothesis pairings. Our first step is to figure out how we can scale this, ensuring that we are training on the entire dataset. We will then work to fine-tune the DeBERTa model to maximize accuracy while simultaneously exploring alternative models in the NLI space. Our final step will be to build an interactive interface using Flask for users to input their own premise/hypothesis pairing to get classified in real time.