Classification using Sentence Embeddings Generated using Task Instructions

Nidhi Chowdhry

Notes from Instructor

- This is a template. You don’t have to follow it exactly, but your presentation should cover what are mentioned in the template.
- Time: 5-6 minutes for each team
  - Followed by 1 minute break for Q&A and transition
- **Some tips:**
  - **Please be considerate to your classmates.** Most of them may have never seen the task you plan to work on, so please include the necessary background knowledge for them.
  - **Use figures and animations smartly!** They will save you a lot of effort:)

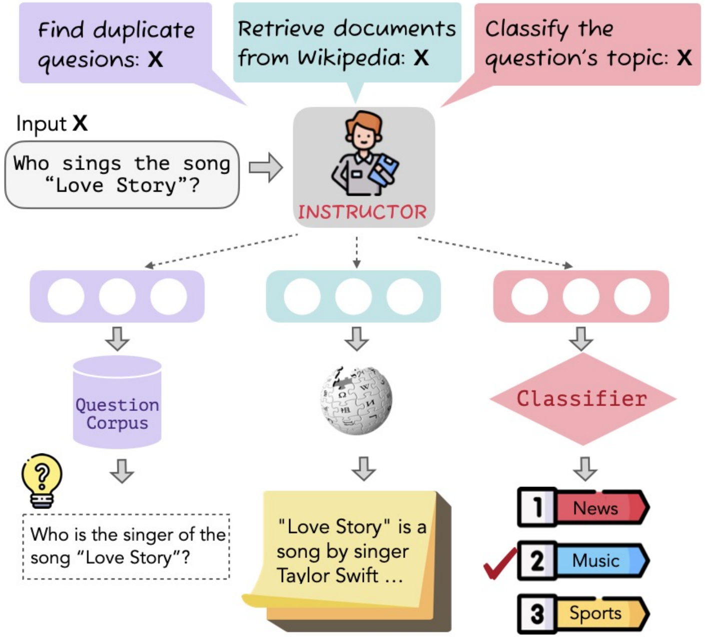
Outline

- Introduction
- Approach
- Experiments
- Conclusion & Future Work

Introduction

Paper - One Embedder, Any Task: Instruction-Finetuned Text Embeddings
Authors - Hongjin Su, Weijia Shi and others

- Sentence transformer
- Existing embeddings can have significantly degraded performance when applied to new tasks or domains
- Instructor - Single multitask model that generates task and domain aware embeddings given a text input and its task instructions
- $E_I(I_x, x) = F(I_x \oplus x)$



Find duplicate quesions: X
Retrieve documents from Wikipedia: X
Classify the question's topic: X
Input X
Who sings the song "Love Story"?
INSTRUCTOR
Question Corpus
Classifier
Who is the singer of the song "Love Story"?
"Love Story" is a song by singer Taylor Swift ...
1 News
2 Music
3 Sports

Introduction

Now, give a clear description of your task: what exact problem are you solving or studying in this project? Sometimes it helps to give a real example from your dataset(s)

- Classification of datasets from 2 different domains using the same model
- Verify Instructor improves classification accuracy over base model.
- Verify if Instructor is robust to noise in data and can accept multi-lingual inputs.
- Instructor uses Logistic Regression for evaluation

Approach

- Update baseline code and create unit tests to evaluate the model
- Train model using different hyperparameters – Seed/Learning Rate/Contrastive Loss temperature
- Evaluate Robustness capabilities -
    - Vocabulary/NER/Typos/Simple Negation using MFT/INV tests.
    - Create MFT templates using masks/Invariance tests by adding perturbations
- Evaluate Multi-linguality –
    - GTR/Instructor is web-trained English only.
    - Train another sentence transformer - use-cmlm-multilingual/distiluse-base-multilingual-cased-v2 – Supports 102 languages
    - Translate training dataset from paper to 5 different languages
    - Translate evaluation datasets to 6 different languages – 1 unknown
    - Verify accuracy on each language.

Experiments

Training Dataset – Use MEDI dataset (specifically created for Instructor).

- MEDI is diverse and combines multiple datasets and add instructions to it. Used combination of Symmetric and Asymmetric Wikipedia (WikiAnswers, WikiHow, simple_wiki) datasets - ~35000 instruction/input pairs

Evaluation Datasets –

- MTEB Classification Datasets
  - mteb/emotion - Emotion is a dataset of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise.
  - Mteb/mtop_domain - Task-Oriented Semantic Parsing. – messaging/calling/event/timer/weather/music/alarm/people/recipes/news

- Evaluation Metric:

Accuracy + F1 Score (dissimilar class distribution)

Results

Baseline Results:

Dataset	GTR-Base	GTR	INSTRUCTOR
emotion	42.2	45.5	53.2
mtopdomain	92.42	93.6	95.1

Checkpoint 1 result
(trained on GTR Base
Average across 2 seeds and 2 LR)

Dataset	LR 2e-4	LR 2e-5
Emotion (s=default)	46.82	49.92
Emotion (s=30)	47.00	49.92
Mtopdomain (s=default)	92.65	92.20
Mtopdomain (s=30)	90.30	92.20

Accuracy
base use-cmlm-multilingual vs
instructor trained use-cmlm-multilingual

Dataset	Base CMLM	LR 2e-4	LR 2e-5
Emotion (s=30)	26.65	29.05	33.52
Mtopdomain (s=30)	89.22	77.74	84.73

Robustness Accuracies

Capabilities	Base Instructor	Trained Instructor	Base CMLM	Trained CMLM	Base Instructor	Trained Instructor	Base CMLM	Trained CMLM
	Emotion Dataset				MTOPDomain Dataset			
Vocabulary	63.33				87.27			
Robustness (Typos)	13.33				5.45			
NER	20.00				100			
Negation	25.00				100			
Reverse Negation	50.00				100			

Future Work

- Complete all Robustness and Multi-linguality tests for final report.
- Error analysis for Robustness

Other areas to explore in future:

- Verify if Instructions improve accuracy for different task types – Clustering/retrieval etc. on various domains.
- Robustness/multilinguality/cross-linguality evaluation of task instructions

Thank you!