NUS | Computing
National University of Singapore

CS4248 Natural Language Processing

Project Dataset Descriptions

Labeled Unreliable News (LUN)

Type: Document Classification, Misinformation Detection

Size: 48K news articles for training, 3K for testing.

Estimated difficulty: **easy (2-way), medium (4-way)**

Compute cost estimate: **low–medium**

Website: [download data]; refer to the .csv files

(dataset originally constructed by Rashkin *et al.* (2017))

Predict the reliability of a news document; Either 4-way (trusted, satire, hoax, propaganda) or 2-way (trusted, satire)

Labels: 1-"Satire", 2-"Hoax", 3-"Propaganda", 4-"Reliable News"

Path on ConceptNet

Type: Classification (can be adapted for a Generation task)

Size: 20K

Estimated difficulty: **medium**

Compute cost estimate: **low**

Website: https://github.com/YilunZhou/path-naturalness-prediction

Predict the naturalness of a path within ConceptNet.

Zhou, Schockaert and Shah (2019)

SciCite: Citation Intent Classification

Type: Scientific Document Processing, Sentiment Analysis, Sentence Classification

Size: 11K

Estimated difficulty: **medium**

Compute cost estimate: **medium**

Website: https://github.com/allenai/scicite

Given an input citation sentence (“context”), classify its sentiment / intent as one among {background, method, comparison}

IWSLT 2017, Chinese–English

Type: Translation (Sequence Generation)

Size: 230K paired sentences for training, 8.5K for testing

Estimated difficulty: **medium**

Compute cost estimate: **high**

Website: HuggingFace Repository

Reference: IWSLT 2017 Datasets

Paired subscripts of TED talks. This dataset is suitable for building sentence-level translation systems, for both English–Chinese and Chinese–English directions.

e-SNLI

Type: Classification / Generation

Size: 570K

Estimated difficulty: **medium**

Compute cost estimate: **very high**

Website: https://github.com/OanaMariaCamburu/e-SNLI

Builds on top of Stanford Natural Language Inference (SNLI) dataset.

Explanation task: Given a premise and a hypothesis, generate an explanation for a predicted label.

Prediction task: Given a tuple of a premise, hypothesis and the explanation, make a prediction.

WI-LOCNESS

Type: Generation

Size: 43k

Estimated difficulty: **medium**

Compute cost estimate: **medium**

Website: https://www.cl.cam.ac.uk/research/nl/bea2019st/

A grammatical error correction dataset. The source sentence is a sentence with or without grammar errors, and the target sentence is a sentence without any grammar errors.

GSM8K

Type: Reasoning (Generation)

Size: 7.5K

Estimated difficulty: **very high**

Compute cost estimate: **very high**

Website: https://github.com/openai/grade-school-math

It is a high quality linguistically diverse grade school math word problem.