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"

ELCo-AN Emoji Composition

Type: Emoji, Lexical Compositionality, Ranking,

Size: 210 adjective–noun compounds, multiply annotated ~7 times

Estimated difficulty: **easy, medium, hard (challenge for publication)**

Compute cost estimate: **low–medium**

A dataset consisting of emoji representations of adjective–noun compounds (AN). Unpublished NUS research. Internal documentation to be provided soon by CS4248 teaching staff before Week 04.

Rank the plausibility of an English phrase to explain an emoji combination:

- ❄️🐻 = Polar Bear (high), Office Worker (low)

*If done well, your team may be able to participate in co-publication of this research work with the current team led by TA Yisong.

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)

Dialogue Relation Extraction

Type: Information Extraction, Sequence Classification

Size: 1.7K dialogues. Each dialogue includes multiple pairs of entities.

Estimated difficulty: **medium**

Compute cost estimate: **medium**

Website: Yu *et al.* (2020) (The data is the paper’s GitHub repo)

Given a dialogue as context, and a few entity pairs mentioned in the dialogue, this task is to classify the relation type of each pair of entities. The relation type is pre-defined.

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.

SQuAD: The Stanford Question Answering Dataset

Type: Token classification

Size: 100K + questions

Estimated difficulty: **medium–high**

Compute cost estimate: **medium– high**

Website: https://rajpurkar.github.io/SQuAD-explorer/

Note: please use the v1.1 dataset as Version 2 adds extra complexity

A reading comprehension dataset containing questions and corresponding context paragraphs. Given a question, the task is to answer the question by extracting tokens from its context paragraph.

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.