Overview

- Construct the first NLU benchmark in Japanese.
- Tasks are chosen to cover GLUE [Wang+ 18] and SuperGLUE [Wang+ 19] tasks.

Introduction

- To develop high-performance NLU models, a benchmark is necessary.
- In the case of English, the GLUE Benchmark is publicly available.
- While benchmarks for languages other than English have been constructed, there is no benchmark for Japanese.
  - Chinese: CLUE [Xu+ 19], French: FLUE [Le+ 20], etc.

Construction of Japanese Benchmark

- Due to the linguistic characteristics of Japanese, findings in other languages cannot necessarily be applied.
  - The Japanese alphabet includes hiragana, katakana, Chinese characters, and the Latin alphabet.
  - There are no spaces between words. (I had two pieces of bread this morning.)
- Problems with existing Japanese datasets
  - Translation (e.g., JSNLI [Yoshikoshi+ 20])
  - Unnaturalness of Japanese in machine/manual translation
  - Specific domains (e.g., JRTSE [Hayashibe+ 20])
  - Not suitable for evaluating NLU ability in general domain.

Build a Japanese Language Understanding Benchmark and facilitate NLU research for Japanese.

Datasets in JGLUE

- **MARC – ja**
  - A sentiment classification task (positive or negative) for product reviews.
  - Build on the Japanese part of MARC (Multilingual Amazon Reviews Corpus) [Keung+ 20].
  - Label modification through a crowdsourced label judgment task.

- **JSQuAD**
  - A Japanese version of SQuAD [Rajpurkar+ 16], a general domain QA dataset constructed using Wikipedia.
  - Our construction method is based on SQuAD.

- **JCommonsenseQA**
  - Five choice QA to evaluate commonsense reasoning ability.
  - Our construction method basically follows CommonsenseQA [Talmor+ 19].
  - Remove QSS that contain synonyms in targets to make the answer unique.

Evaluation using JGLUE

- Evaluate the performance of pre-trained models using JGLUE.

Findings

- Overall, XLM-RoBERTa-lg-head performed the best.
  - 1. LARGE model size, 2. CommonCrawl > Wikipedia
  - Tokenization: subword base > character base
  - Models pre-trained using CommonCrawl performed well.

Datasets in JGLUE

- **JSQuAD**
  - **JSTS**
    - A task of predicting the semantic similarity between two sentences with a value between 0 (completely different in meaning) and 5 (equivalent in meaning).
  - **JNLI**
    - A task of recognizing the inference relation that a premise sentence has to a hypothesis sentence based on 3 labels (entailment, neutral, and contradiction).

We basically extract sentence pairs in JSTS and JNLI from YJ Captions Dataset [Miyazaki+ 16].