JGLUE: Japanese General Language Understanding Evaluation

Kentaro Kurihara[1], Daisuke Kawahara[1], Tomohide Shibata[2]
Waseda University[1], Yahoo Japan Corporation[2]

Overview

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

Task	Dataset	Train	Valid	Test
Text Classification	MARC – ja	187,528	5,654	5,639
	JCoLA [Someya+ 22]	---	---	---
Sentence Pair Classification	JSTS	12,451	1,457	1,589
	JNLI	20,073	2,434	2,508
QA	JSQuAD	63,870	4,475	4,470
	JCommonsenseQA	9,012	1,119	1,118

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.

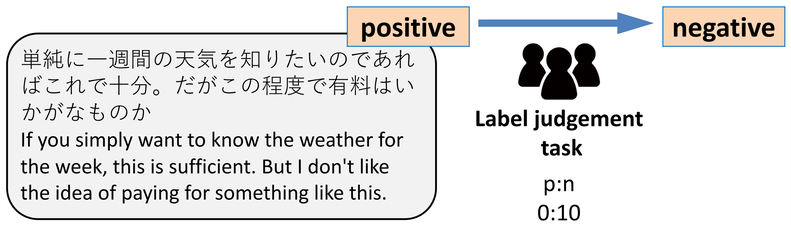
    私は今朝パンを2枚食べました。
    (I had two pieces of bread this morning.)
- Problems with existing Japanese datasets
  - Translation (e.g., JSNLI [Yoshikoshi+ 20]) — From Scratch
    - Unnaturalness of Japanese in machine / manual translation
  - Specific domains (e.g., JRTE [Hayashibe+ 20]) — In General Domain
    - 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 judgement task.



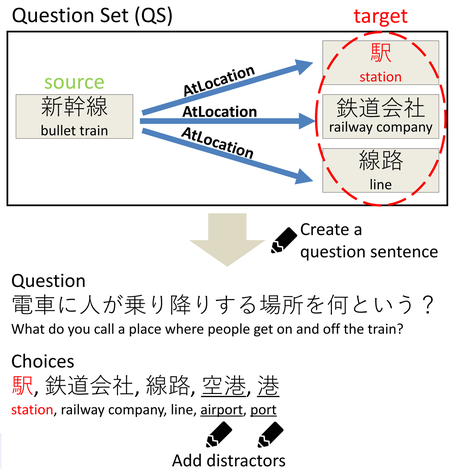
JSQuAD

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

JCommomsenseQA

- 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.

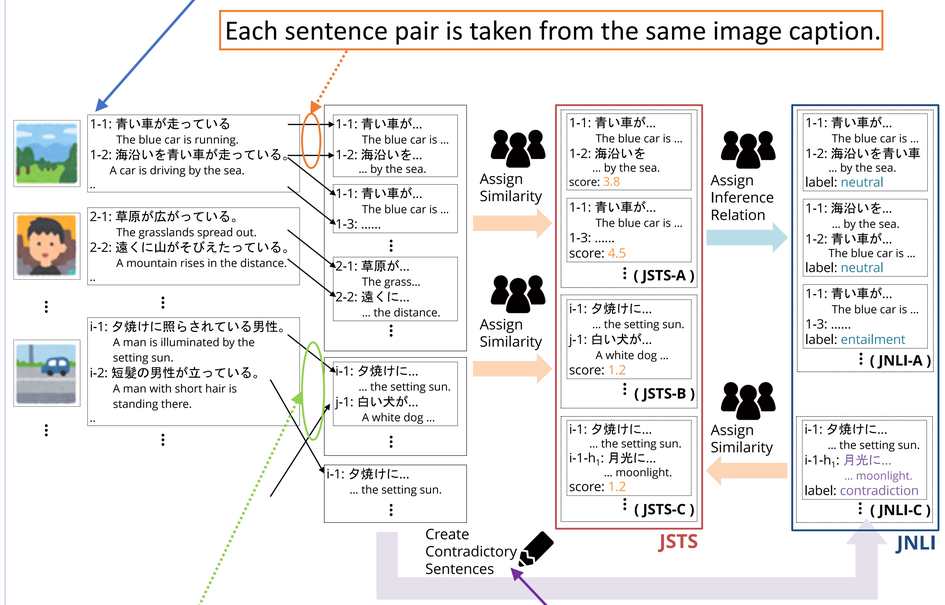
庭 (garden) ≒ 庭園 (garden)



JSTS / JNLI

- 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].



Evaluation using JGLUE

- Evaluate the performance of pre-trained models using JGLUE.
- Findings
  - Overall, XLM-RoBERTa$_{LARGE}$ performed the best.
    - 1. LARGE model size, 2. CommonCrawl > Wikipedia
  - Tokenization: subword base > character base
  - Models pre-trained usging CommonCrawl performed well.

	MARC-ja	JSTS	JNLI	JSQuAD	JComQA
Models	acc	Pearson	acc	F1	acc
Human	0.990	0.909	0.917	0.946	0.988
Tohoku BERT$_{BASE}$	0.957	0.901	0.876	0.946	0.782
Tohoku BERT$_{BASE}$ (char)	0.957	0.889	0.861	0.937	0.728
Tohoku BERT$_{LARGE}$	0.961	0.907	0.878	0.950	0.822
NICT BERT$_{BASE}$	0.960	0.909	0.881	**0.952**	0.807
Waseda RoBERTa$_{BASE}$	0.962	0.901	0.876	0.926	**0.849**
XLM-RoBERTa$_{LARGE}$	**0.965**	**0.916**	**0.902**	—	0.842

Conclusion / Future work

- Constructed JGLUE, the first NLU benchmark in Japanese.
- JGLUE is available at https://randd.yahoo.co.jp/en/softwaredata#jglue
- Plan to build Japanese benchmarks for generation tasks such as GLGE [Liu+ 21].