Multi-Lingual Natural Language Processing

Homeworks Report

SAPIENZA
UNIVERSITÀ DI ROMA

Dario Loi — 1940849

M.Sc. in AI & Robotics,
Sapienza, University of Rome.

A. Y. 2023–2024

Tasks

For the first homework, I was assigned two tasks:

1. WiC-ITA: Detect whether two italian words in two different sentences are used with the same meaning.
2. ITAmoji: Predict which emoji was used in a given italian tweet.

We will spend a few slides on each task.

WiC-ITA

The first task involved the parsing of JSONL files containing the Word-in-Context-ITA dataset, an example of which is shown below:

Listing: Sample 23 from WiC-ITA `train.jsonl`.

```
{
  "id": "lira.noun.15",
  "lemma": "lira",
  "sentence1": "In caso di inosservanza degli obblighi stabiliti dal comma 1 , si applica la
       sanzione amministrativa pecuniaria da lire dieci milioni a lire cento milioni .",
  "sentence2": "Per le finalità di cui all' Art. 1 , comma 2 , della legge regionale 26 aprile 1
       995 , n. 31 recante \" Norme in materia di musei degli Enti locali e di interesse locale
        \" è autorizzata per l' esercizio finanziario 2000 la spesa di lire 200.000.000 .",
  "start1": 115,
  "end1": 119,
  "start2": 230,
  "end2": 234,
  "label": 1
}
```

Desired Format

The samples were rewritten for use by an LLM, as shown below:

Listing: Same sample with required changes.

```
{
  "id": "lira.noun.15",
  "lemma": "lira",
  "sentence1": "In caso di inosservanza degli obblighi stabiliti dal comma 1 , si applica la
       sanzione amministrativa pecuniaria da lire dieci milioni a lire cento milioni .",
  "sentence2": "Per le finalità di cui all' Art. 1 , comma 2 , della legge regionale 26 aprile
       1995 , n. 31 recante \" Norme in materia di musei degli Enti locali e di interesse
      locale \" è autorizzata per l' esercizio finanziario 2000 la spesa di lire 200.000.000
       .",
  "start1": 115,
  "end1": 119,
  "start2": 230,
  "end2": 234,
  "choices": [
    "DIVERSO",
    "UGUALE"
  ],
  "label": 1
}
```

ITAmoji

The second task also involved manipulation of JSONL files, a typical sample of ITAmoji is shown below:

Listing: Sample 27 from `ITAmoji_2018_TRAINdataset_v1.ANON.list`.

```
{
  "uid": "447352763",
  "text_no_emoji": "... il rumore del mare \ufe0f #28Settembre <URL>",
  "created_at": "Thu Sep 28 15:32:06 +0000 2017",
  "label": "red_heart",
  "tid": "913426094002458626"
}
```

For this task, we also have to add distractors, that is, plausible alternatives to the correct label.

Desired Format... Again

The output from our distractor generation process on the previous sample is as follows:

Listing: Sample 27 with generated distractors.

```
{
  "id": "ITA-emoji-train-00000027",
  "sentence": "... il rumore del mare \ufe0f #28Settembre <URL>",
  "choices": [
      "red_heart",
      "two_hearts",
      "rose",
      "kiss_mark"
  ],
  "label": 0
}
```

Distractor Clusters

To generate distractors, we used a hand-crafted list of semantic clusters, which we define as a set of emoji that are semantically related. For example, the cluster `love` contains the emojis `red_heart`, `two_hearts`, `blue_heart`, `rose`, and `kiss_mark`. When augmenting a sample, we select a cluster in which the correct label is present, and then randomly select three other labels from the same cluster. This way, we ensure that the distractors are capable to confuse the model.

Distractor Sampling

To obtain a set of distractors from an ITAmoji sample, we follow this process:

- Randomly sample a cluster from the list of clusters that contain the correct emoji.
- Randomly sample three other emojis from the same cluster.
- Return the new sample with the correct label and the three distractors.

Distractor Sampling

To obtain a set of distractors from an ITAmoji sample, we follow this process:

- Randomly sample a cluster from the list of clusters that contain the correct emoji.
- Randomly sample three other emojis from the same cluster.
- Return the new sample with the correct label and the three distractors.

Distractor Sampling

To obtain a set of distractors from an ITAmoji sample, we follow this process:

- Randomly sample a cluster from the list of clusters that contain the correct emoji.
- Randomly sample three other emojis from the same cluster.
- Return the new sample with the correct label and the three distractors.

Distractor Sampling

To obtain a set of distractors from an ITAmoji sample, we follow this process:

- Randomly sample a cluster from the list of clusters that contain the correct emoji.
- Randomly sample three other emojis from the same cluster.
- Return the new sample with the correct label and the three distractors.