NLP Course Report

Hans Peter Lyngsøe, pvr448

September 5, 2024

1 Week 1

(a) Basic statistics:

- the data is quite evenly distributed across the 3 languages.
- We note that there are more answerable than unanswerable by a factor of 10-1.
- train_set: 15326
- val_set: 3028

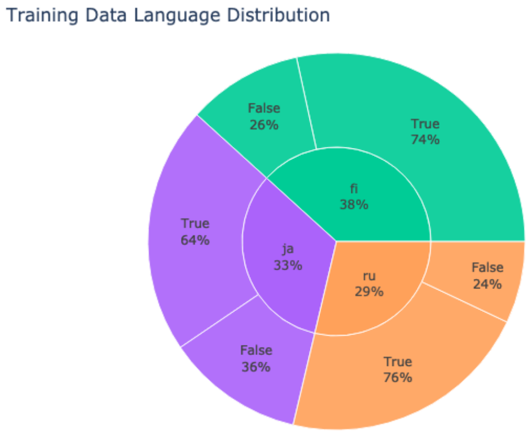

Figure 1: Distribution of labels in the dataset

(b)

```
import MeCab
def get_top_words(df: pd.DataFrame, lang: str,
    n=5):
    df_lang = df[df[’lang’] == lang].copy()
```

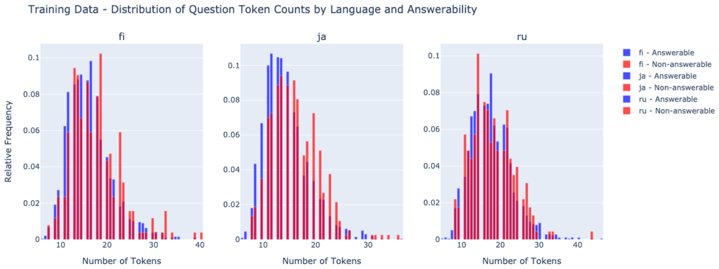

Figure 2: normalized Histogram for token count(using llama3 tokenizer) of answerable/unanswerable questions in the dataset

```
if lang == ’ja’:
    mecab = MeCab.Tagger("-Owakati")  #
        Initialize MeCab tokenizer
    df_lang.loc[:, ’words_question_tokens’
        ] = df_lang[’question’].apply(
        lambda x: mecab.parse(x).split())
else:
    df_lang.loc[:, ’words_question_tokens’
        ] = df_lang[’question’].apply(
        lambda x: x.split(’␣’))

all_tokens = np.concatenate(df_lang[’
    words_question_tokens’].values)
unique, counts = np.unique(all_tokens,
    return_counts=True)
sorted_indices = np.argsort(counts)[::-1]
top_unique_tokens = unique[sorted_indices
    ][:n]
top_tokens_dict = {token: int(count) for
    token, count in zip(top_unique_tokens,
    counts[sorted_indices][:n])}
return top_tokens_dict
```

we get the following results:

(c)

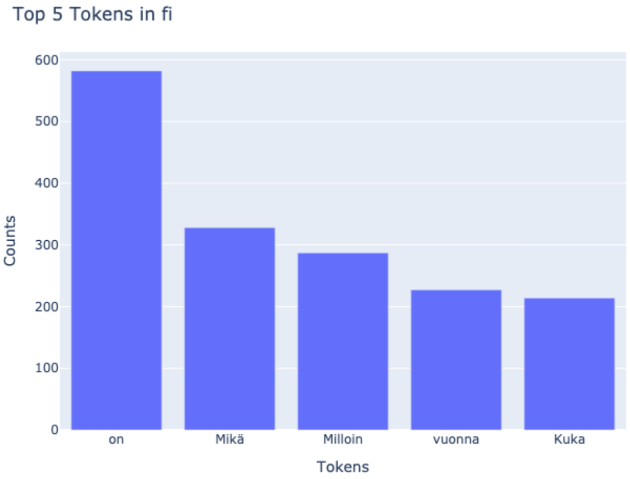

Figure 3: Top 5 tokens in Finnish

2 Week 37 (9–15 September)

3 Week 38 (16–22 September)

4 Week 39 (23–29 September)

5 Week 40 (30 September–6 October)

6 Week 41+ (from 7 October)

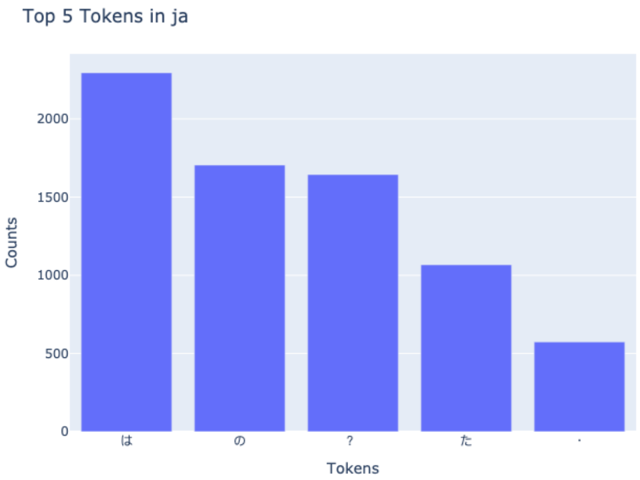

Figure 4: Top 5 tokens in Japanese

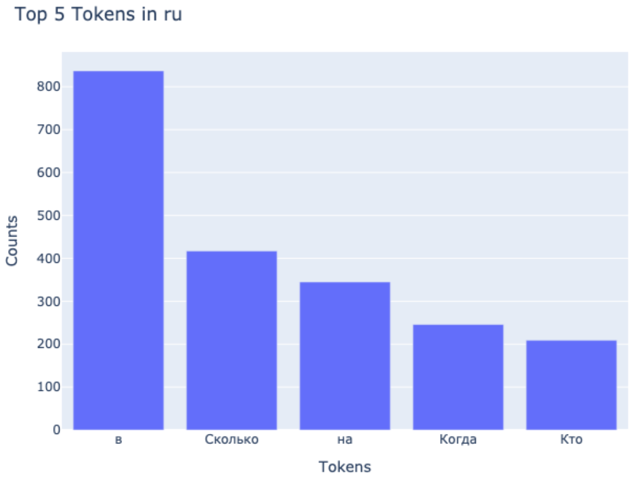

Figure 5: Top 5 tokens in Russian