**Prof. Dr. Goran Glavaš,**
**M.Sc. Fabian David Schmidt**
**M.Sc. Benedikt Ebing**
Lecture Chair XII for Natural Language Processing, Universität Würzburg

# 7. Exercise for "Multilingual Natural Language Processing"

12.07.2024

## 1 Paper Readings

We segment the literature on neural machine translation (NMT) as follows:

1. **Bitext Mining**
   - CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web
2. **NMT with Large Language Models**
   - Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis

## 2 Bitext Mining

Modern supervised NMT systems like NLLB achieve remarkable translation performance even on low-resource languages. In this part of the exercise, we will explore one key building block of such NMTs called bitext mining.

1. Briefly explain bitext mining.

> Automatically retrieving parallel sentences (i.e., sentences that are translations of each other) from large corpora (e.g., the web).

2. Why is bitext mining important for NMT? Explain the motivation.

> Parallel sentences are crucial for the training of multilingual sentence encoders and neural machine translation. For higher resource languages, comparably large parallel corpora exist (e.g., United Nations Parallel Corpus or Europarl). However, specifically for lower resource languages, there are less efforts to manually construct parallel corpora, creating the need for automatically mined parallel data.

3. Summarize the bitext mining method presented in *CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web (Schwenk et al., 2021)*.

> a) Text extraction: Extracting text form the CCNet JSON, deduplicate the sentences, perform language identification
>
> b) Get all sentence embeddings using a sentence encoder (storing them in a compressed from using FAISS library)
>
> c) For all pairs of sentences of two languages, compute the margin-based criterion for both directions (sentence embeddings that are "close" in their respective neighborhood). Built the union of the forward and backward direction, sort the candidates and omit source or target sentences that are already used. Then apply a threshold (hyperparameter) to decide whether two sentences are mutual translations.

# 3 NMT with Large Language Models

Large language models like ChatGPT, Bloomz, or mT0 gain increasing attention as "generalists", i.e., they are able to solve tasks with no or few examples seen. In this section, we examine this hypothesis and investigate whether large language models outperform supervised NMT systems in automatic translation.

1. Do large language models outperform supervised NMT models? Briefly summarize the main results from the paper.

> Compare eight large language models, four that are "only" pretrained (XGLM, OPT, Falcon-7B, LLaMA2-7B) and four instruction-tuned (BLOOMZ, LLaMA2-7B-chat, ChatGPT, GPT-4) against three translation models, two open-source (M2M-100-12B and NLLB-1.3) and the closed-source Google Translate. GPT-4 outperforms NLLB in $40.91\%$ of the tested translation directions (GPT-4 performs best among the large language models), but underperforms Google

Translate (particularly on low-resource languages and translating from English). Open source LLMs still underperform open source NMT. All in all, supervised NMT models still outperform large language models in lower resource languages and translations from English to the target language. On high-resource languages and translating to English close-source LLMs are already competitive.
Remarks:

- Instruction-tuned models are typically fine-tuned on translation tasks
- NLLB-1.3B is not the largest NMT model from that family

2. Large language models are prone to produce certain translation errors. Name and describe the three typical translation errors presented in the paper.

- *Off-target translations*: Translating into the wrong target language (e.g., into Azerbaijani instead of Turkish).
- *Hallucination*: Generating highly pathological translations that are unrelated with the source sentence.
- *Monotonic translations*: Translating sentences word-by-word lacking effective word-reordering of the target language.

3. Describe the issue of "data leakage" when evaluating large language models on publicly available datasets.

Large language (and instruction-tuned) models might be trained on publicly available evaluation data. This effect might inflate the performance on certain evaluation datasets. It is particularly an issue for commercial models that do not open-source their training corpora which hinders fair comparison.

4. How important is the choice of the template for prompting? Briefly describe the corresponding results from the paper.

The performance of large language models relies on the template choice. There is a gap in the average performance of up to 16 BLEU between the best and the worst performing template. However, even unreasonable templates may produce decent translations (e.g., instruct the model to summarize). The role of the template is still an open research area.

# 4 Additional Exercises (Not required for bonus)

1. We are given the following probabilities for tokens {BOS, A, B, C, D} at timesteps $\{T\}_{i=0}^{3}$ for a text generation model.

| | T = 0 | T = 1 | T = 2 | T = 3 |
|---|---|---|---|---|
| BOS | 0.140 | 0.257 | 0.248 | 0.149 |
| A | 0.391 | 0.096 | 0.402 | 0.336 |
| B | 0.197 | 0.341 | 0.267 | 0.358 |
| C | 0.271 | 0.305 | 0.083 | 0.157 |

Compute the probabilities for the decoded sequence for both (i) greedy decoding and (ii) beam search with width $k = 3$. Show your intermediate steps!

**Greedy Decoding**

The decoded string is ABAB, since, at every time step, the corresponding character has the highest probability to be decoded.

$$P(ABAB) = 0.391 \times 0.341 \times 0.402 \times 0.358 = 0.0191 \tag{1}$$

**Beam-Search Decoding**

Use the logarithm of probabilites to avoid underflow.

**Step 1**

| | |
|---|---|
| A | -0.939 |
| C | -1.305 |
| B | -2.297 |

**Step 2**

Intermediate scores computation

| | BOS | A | B | C |
|---|---|---|---|---|
| A | -2.2977 | -3.2825 | -2.0149 | -2.1265 |
| C | -2.6643 | -3.6490 | -2.3815 | -2.4931 |
| B | -2.9832 | -3.9680 | -2.7004 | -2.8120 |

We hence select the next top paths from the above table:

| | |
|---|---|
| A $\rightarrow$ B | -2.0149 |
| A $\rightarrow$ C | -2.1265 |
| A $\rightarrow$ BOS | -2.2977 |

**Step 3**

| | BOS | A | B | C |
|---|---|---|---|---|
| A $\rightarrow$ B | -3.4092 | -2.9262 | -3.3354 | -4.5038 |
| A $\rightarrow$ C | -3.5208 | -3.0378 | -3.4470 | -4.6154 |
| A $\rightarrow$ BOS | -3.6921 | -3.2090 | -3.6182 | -4.7866 |

We hence select the next top paths from the above table:

| | |
|---|---|
| A $\rightarrow$ B $\rightarrow$ A | -2.9262 |
| A $\rightarrow$ C $\rightarrow$ A | -3.0378 |
| A $\rightarrow$ BOS $\rightarrow$ A | -3.2090 |

**Step 4**

| | BOS | A | B | C |
|---|---|---|---|---|
| A $\rightarrow$ B $\rightarrow$ A | -4.8300 | -4.0169 | -3.9534 | -4.7777 |
| A $\rightarrow$ C $\rightarrow$ A | -4.9416 | -4.1284 | -4.0650 | -4.8893 |
| A $\rightarrow$ BOS $\rightarrow$ A | -5.1128 | -4.2997 | -4.2363 | -5.0605 |

We hence select the next top paths from the above table:

| | |
|---|---|
| A $\rightarrow$ B $\rightarrow$ A $\rightarrow$ B | -3.9534 |
| A $\rightarrow$ B $\rightarrow$ A $\rightarrow$ A | -4.0169 |
| A $\rightarrow$ C $\rightarrow$ A $\rightarrow$ B | -4.0650 |

2. Compute the BLEU score step-by-step as per the lecture for the below reference-hypothesis sentence pair.

| | |
|---|---|
| Reference | The quick brown fox jumps over the lazy dog. |
| Hypothesis | The fast brown fox leaps over the lazy dog. |

# BLEU-4 Score Calculation

## Reference and Hypothesis

| | |
|---|---|
| Reference | The quick brown fox jumps over the lazy dog. |
| Hypothesis | The fast brown fox leaps over the lazy dog. |

## Tokenization

- **Reference tokens:** [The, quick, brown, fox, jumps, over, the, lazy, dog]
- **Hypothesis tokens:** [The, fast, brown, fox, leaps, over, the, lazy, dog]

## N-gram Precision Calculation

### Unigrams

- Matching unigrams: [The, brown, fox, over, the, lazy, dog]
- Precision: $p_1 = \frac{7}{9}$

### Bigrams

- Reference bigrams: [(The, quick), (quick, brown), (brown, fox), (fox, jumps), (jumps, over), (over, the), (the, lazy), (lazy, dog)]
- Hypothesis bigrams: [(The, fast), (fast, brown), (brown, fox), (fox, leaps), (leaps, over), (over, the), (the, lazy), (lazy, dog)]
- Matching bigrams: [(brown, fox), (over, the), (the, lazy), (lazy, dog)]
- Precision: $p_2 = \frac{4}{8} = \frac{1}{2}$

### Trigrams

- Reference trigrams: [(The, quick, brown), (quick, brown, fox), (brown, fox, jumps), (fox, jumps, over), (jumps, over, the), (over, the, lazy), (the, lazy, dog)]
- Hypothesis trigrams: [(The, fast, brown), (fast, brown, fox), (brown, fox, leaps), (fox, leaps, over), (leaps, over, the), (over, the, lazy), (the, lazy, dog)]
- Matching trigrams: [(over, the, lazy), (the, lazy, dog)]
- Precision: $p_3 = \frac{2}{7}$

### 4-grams

- Reference 4-grams: [(The, quick, brown, fox), (quick, brown, fox, jumps), (brown, fox, jumps, over), (fox, jumps, over, the), (jumps, over, the, lazy), (over, the, lazy, dog)]
- Hypothesis 4-grams: [(The, fast, brown, fox), (fast, brown, fox, leaps), (brown, fox, leaps, over), (fox, leaps, over, the), (leaps, over, the, lazy), (over, the, lazy, dog)]
- Matching 4-grams: [(over, the, lazy, dog)]
- Precision: $p_4 = \frac{1}{6}$

## Geometric Mean of Precisions

$$\text{Geometric mean} = \exp\left(\frac{1}{4}\sum_{n=1}^{4}\log p_n\right)$$

$$\sum_{n=1}^{4}\log p_n = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

$$\log p_1 = \log\left(\frac{7}{9}\right), \quad \log p_2 = \log\left(\frac{1}{2}\right), \quad \log p_3 = \log\left(\frac{2}{7}\right), \quad \log p_4 = \log\left(\frac{1}{6}\right)$$

$$\sum_{n=1}^{4}\log p_n = \log\left(\frac{7}{9}\right) + \log\left(\frac{1}{2}\right) + \log\left(\frac{2}{7}\right) + \log\left(\frac{1}{6}\right)$$

$$\sum_{n=1}^{4} \log p_n = \log\left(\frac{7}{9} \times \frac{1}{2} \times \frac{2}{7} \times \frac{1}{6}\right)$$

$$\sum_{n=1}^{4} \log p_n = \log\left(\frac{1}{54}\right)$$

$$\text{Geometric mean} = \exp\left(\frac{1}{4}\log\left(\frac{1}{54}\right)\right) = \exp\left(\log\left(\left(\frac{1}{54}\right)^{1/4}\right)\right) = \left(\frac{1}{54}\right)^{1/4}$$

$$\left(\frac{1}{54}\right)^{1/4} = \frac{1}{54^{1/4}}$$

## Brevity Penalty (BP)

- Reference length $r = 9$
- Hypothesis length $c = 9$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases}$$

Since $c = r$:

$$\text{BP} = 1$$

## Final BLEU-4 Score Calculation

$$\text{BLEU-4} = \text{BP} \times \text{Geometric mean} = 1 \times \left(\frac{1}{54}\right)^{1/4}$$

Simplifying:

$$\text{BLEU-4} = \left(\frac{1}{54}\right)^{1/4} \approx 0.3688$$