UDC 004.89:004.93

EVALUATING LANGUAGE MODELS ON LOW-RESOURCE PAIRS

Bodenchuk-Pastukhov Y. V.
email: yehor.bodenchuk-pastukhov@nure.ua
Supervisor – Candidate of Technical Sciences, Assistant Kobylin I. O.
Kharkiv National University of Radio Electronics, dep. INF
c. Kharkiv, Ukraine

This work is devoted to the evaluation of the Facebook M2M100_418M and Alirezamsh Small100 models for low-resource language pairs. For this study, parallel corpora were selected for the following language pairs: Japanese-Ukrainian, Korean-Ukrainian, Turkish-Ukrainian, Vietnamese-Ukrainian, and Chinese-Ukrainian. The models were assessed based on their performance in translating these language pairs. Evaluation metrics included BLEU and ChrF scores, which measure the quality of the translations. Additionally, differences between the target and translated sentences were analyzed. The study aims to highlight the strengths and weaknesses of each model when working with low-resource languages. A comparative analysis of the results provides insights into the effectiveness of these models. The findings can be useful for future improvements in machine translation for underrepresented language pairs.

First of all, the open-source project Tatoeba was used to obtain datasets for evaluating Japanese-Ukrainian, Korean-Ukrainian, Turkish-Ukrainian, Vietnamese-Ukrainian, and Chinese-Ukrainian parallel corpora. Each dataset contains 1,000 to 2,000 sentence pairs, with source and target language translations. After running the evaluation process, we obtained the following examples of translations.

Table 1 – Example of translation file from Japanese to Ukrainian

Source	Target	Translated
本当？	Справді?	Насправді ?
雄弁は銀、沈黙は金。	Слово - срібло, мовчання - золото.	Слово – це срібло, мовчання – це золото.
愉快な夏休みでありますように！	Бажаю вам гарного літнього відпочинку!	Нехай у вас буде веселий літній відпочинок!
野菜の値段が下がっている。	Ціни на овочі знижуються.	Ціни на овочі знижуються.

For evaluating we will use BLEU and ChrF for getting metrics. BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine’s output and

that of a human: «the closer a machine translation is to a professional human translation, the better it is» – this is the central idea behind [1]. BLEU can be calculated as follows (1):

$$BLEU = BP \times exp(\sum_{n=1}^{N} w_n log p_n) \quad (1)$$

where *BP* is brevity penalty, which penalizes translations that are too short, p_n is *n*-gram precisions, measuring the overlap of *n*-grams between the generated and reference translations, w_n is weight assigned to each *n*-gram precision term, N is maximum order of *n*-grams considered.

ChrF (Character *F*-score) is a machine translation evaluation metric that operates at the character level instead of the word level, making it particularly useful for morphologically rich languages [2]. It is based on the *F*-score using precision and recall of *n*-grams at the character level. ChrF can be calculated as follows (2):

$$ChrF = (1+\beta^2) \times \frac{((\frac{1}{N}\sum_{(k=1)}^{N} P_k) \times (\frac{1}{N}\sum_{(k=1)}^{N} R_k))}{(\beta^2 \times ((\frac{1}{N}\sum_{(k=1)}^{N} P_k) \times (\frac{1}{N}\sum_{(k=1)}^{N} R_k))} \quad (2)$$

where N is the maximum n-gram order, β is weighting parameter (typically set to 2 to favor recall slightly more than precision), P_k and R_k are precision and recall for *n*-grams of length k.

In the result we have such graphic of comparing BLEU scores:

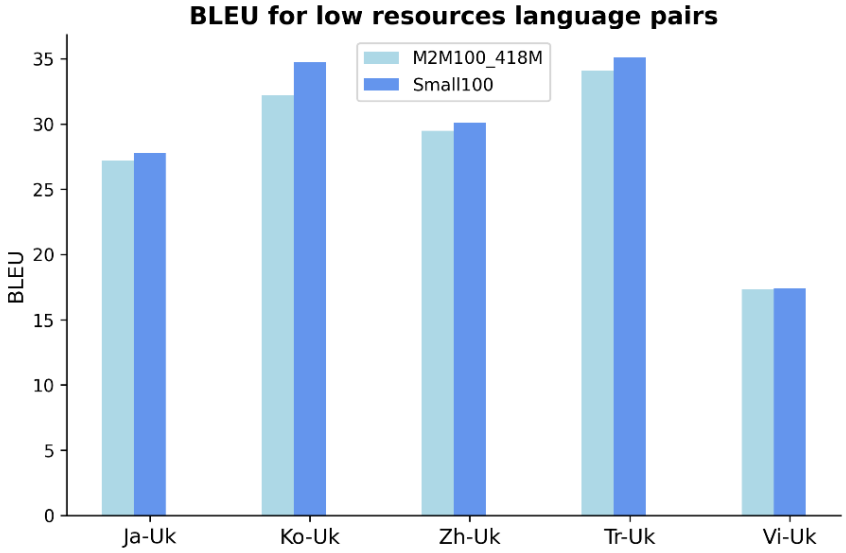

Fig. 1 – Graphic of comparing BLEU score of research models

Also images of difference length of target and translated sentences were configured below. The closer dot to 0 the better is translation to origin.

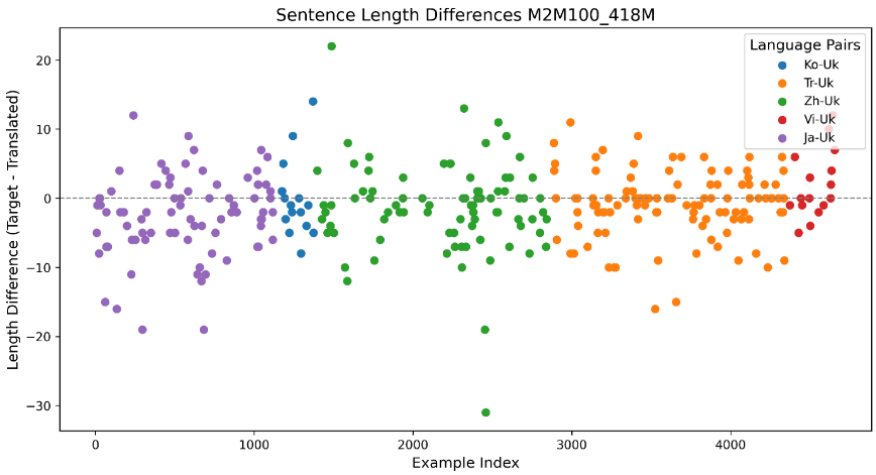

Fig. 2 – Difference of target and translated sentences of M2M100_418M model

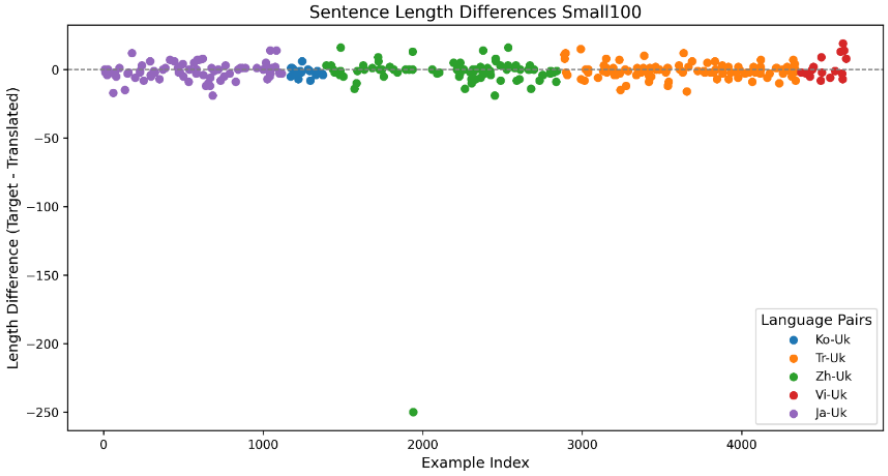

Fig. 3 – Difference of target and translated sentences of Small100 model

Based on all evaluation metrics, the Small100 model outperforms M2M100_418M across all tested low-resource language pairs. It demonstrates higher BLEU and ChrF scores, as well as better alignment between source and target translations. Given its superior performance, Small100 is a more suitable candidate for fine-tuning on specific low-resource pairs. Further research should focus on optimizing its training process and adapting it to domain-specific datasets to enhance translation quality even further.

References:

1. Hugging face – metric BLEU: website. URL: https://huggingface.co /spaces/evaluate-metric/bleu/ (date of application: 01.03.2025)

2. Hugging face – metric ChrF: website. URL: https://huggingface. co/spaces/evaluate-metric/chrf (date of application: 01.03.2025)