Показатели качества работы переводчиков играют важную роль в оценке и сравнении качества машинного перевода. Они используются для определения степени точности и естественности перевода, выполняемого автоматическими системами. Далее мы адаптируем метрики качества перевода к проблеме поиска решений дифференциальных уравнений. Ниже рассмотрены наиболее популярные метрики. Метрика BLEU (Bilingual Evaluation Understudy) [1] - одна из самых известных метрик качества машинного перевода. Она была предложена в 2002 году и стала стандартом для оценки качества машинного перевода. BLEU измеряет качество перевода, сравнивая его с одним или несколькими эталонными переводами, выполненными профессиональными переводчиками. BLEU рассчитывает точность n-грамм (последовательностей из n слов) в машинном переводе по отношению к референсным переводам, а затем усредняет результаты по всем n-граммам (обычно от 1 до 4). Значение BLEU варьируется от 0 до 1, где 1 означает идеальное совпадение с переводом-референсом. Однако достичь значения 1 в реальных условиях практически невозможно. При использовании метрик качества перевода для нашей задачи в качестве эталонного перевода используется решение дифференциального уравнения, взятое либо из датасета, построенного на основе учебников, либо из сгенерированного датасета. Затем решение уравнения, построенное большой языковой моделью, сравнивается с этим эталонным решением.

Помимо метрики BLEU, часто используются и другие метрики: METEOR (Metric for Evaluation of Translation with Explicit Ordering) [2], TER (Translation Edit Rate) [3] и CHRF (Character n-gram F-score) [4]. METEOR учитывает совпадение слов между машинным и эталонным переводом, расстояние между совпадающими словами, а также синонимы и морфологические варианты слов. TER измеряет количество правок (таких как вставка, удаление, замена или перемещение слов), которые необходимо выполнить, чтобы привести машинный перевод в соответствие с эталонным, и показывает долю правок по отношению к общему количеству слов в эталонном переводе. В CHRF используется F-score для сравнения символьных n-грамм между машинным и эталонным переводами. Несмотря на то, что метрика BLEU не всегда точно отражает качество перевода, особенно при работе с идиоматическими выражениями и сложными синтаксическими структурами, она является наиболее популярной, поэтому было решено использовать именно ее в нашей работе. Кроме того, полученные результаты предсказания решений дифференциальных уравнений по метрике BLEU достигает лишь 0.569, то есть, 56.9 баллов, что соответствует уровню: хорошее соответствие. Текст в целом совпадает с эталонным переводом, хотя еще возможно наличие некоторых ошибок.

Ниже приведена градация уровней метрики BLEU:

0 баллов: Нет никакого совпадения между машинным переводом и эталонным переводом.

1-20 баллов: Перевод очень далек от эталонного. Обычно такие значения указывают на плохое соответствие.

20-40 баллов: Среднее качество перевода. Возможны ошибки, и переводы могут не всегда быть точными лексически и синтаксически.

40-60 баллов: Хорошее соответствие. Текст в целом совпадает с эталонным переводом, хотя еще возможно наличие некоторых ошибок.

60-80 баллов: Очень хорошее качество перевода, близкое к эталонному.

80-100 баллов: Почти полное совпадение, что редко встречается в практической оценке, кроме случаев с простыми текстами или очень ограниченными ансамблями.

Важно понимать, что оценка BLEU выше 70-80 баллов считается очень высоким показателем в практике машинного перевода. В реальных условиях достижение 40-60 баллов уже может считаться успехом в зависимости от сложности и разнообразия тестового набора данных.

Ссылки.

1. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, USA (2002).

2. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Workshop on Statistical Machine Translation 2007, Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231. Association for Computational Linguistics, USA (2007).

3. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Conference of the Association for Machine Translation in the Americas 2006, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pp. 223-231. Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA (2006).

4. Popovic, M.: CHRF: Character n-gram F-score for Automatic MT Evaluation. In: Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., Pecina, P. (eds) Workshop on Statistical Machine Translation 2015, Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 392–395. Association for Computational Linguistics, USA (2015).