**Д. А. Маурин**
*Тверской государственный университет, магистрант*
*Научный руководитель: к.ф.н. С. А. Колосов*

# ПРИМЕНИМОСТЬ МЕТРИКИ BLEU ДЛЯ ОЦЕНКИ КАЧЕСТВА МАШИННОГО ПЕРЕВОДА

Цель нашей научно-исследовательской работы заключается в выявлении преимуществ и недостатков метрики BLEU при оценке качества машинного перевода с английского языка на русский и французский на основе научно-популярных текстов волонтерского переводческого проекта Unique.

Актуальность исследования заключается в потребности оптимизировать и автоматизировать оценку качества машинного перевода. Практическая ценность заключается в предоставлении инструмента для сравнительного анализа и улучшения систем машинного перевода (СМП).

BLEU — это показатель качества для систем вывода текста, который пытается измерить соответствие между результатами машинного перевода и человеческим переводом. Основная идея BLEU заключается в том, что чем ближе машинный перевод к профессиональному человеческому переводу, тем он лучше [Ахрамеева 2020: 75].

Метрика основывается на статистическом анализе n-граммных совпадений, что позволяет количественно оценить эквивалентность перевода с точки зрения его формальной близости к эталонному варианту [Papineni, Roukos, Ward, Zhu 2001: 2]. Например:

Эталон: *The cat sits on mat.*

Кандидат: *The cat sits on table.*

Точность униграммов (1-грамм) 4 из 5, коэффициент 0,80.

Чтобы избежать переоценки повторяющихся слов, используется модифицированная точность, где количество совпадений ограничивается максимальным числом вхождений [Митренина, Мухамбеткалиева 2021: 80]. Например:

Эталон: *The cat sits on mat.*

Кандидат: *cat cat cat cat cat.*

Все 5 униграммов совпадают с текстом эталона, но проблема в том, что данная лексема встречается лишь один раз в тексте-референсе, поэтому засчитывается только одно использование слова «cat». Одно из пяти совпадений соответственно дает коэффициент, равный 0,20.

Нередко встречаются случаи, когда лексика перевода совпадает с эталоном, но длина текста отличается от исходного варианта. В этом случае метрика предлагает систему штрафов, которая предотвращает ситуацию, когда системы машинного перевода выдают очень короткие, но пословные переводы, которые не передают полного смысла текста [там же].

Стоит отметить, что метрика никак не оценивает синтаксис предложения, она лишь сравнивает лексическое соответствие. Данный аспект может рассматриваться как преимущество, так и недостаток. Преимущество заключается в том, что метрика никак не «штрафует» за перестановку слов. Недостатком же является то, что BLEU будет выдавать достаточно высокий коэффициент для предложений, в которых присутствует большое количество синтаксических ошибок [Satanjeev, Alon 2005: 72]. Влияние метрики на оценку качества машинного перевода зависит от того, характеризуется ли язык фиксированным или свободным порядком слов. Пример языка с относительно свободным порядком слов (русский):

Эталон: *Кот сидит на коврике*.

Кандидат: *На коврике сидит кот*.

В этом случае метрика выдает коэффициент, равный единице, т.е. полное совпадение. Поскольку русский язык характеризуется нефиксированным порядком слов, данная метрика демонстрирует высокую эффективность при оценке текста-кандидата по сравнению с языками с фиксированным порядком слов. Пример языка с фиксированным порядком слов (английский):

Эталон: *Yesterday, I went to the park and played football.*

Кандидат: *I, the yesterday park to played and football went*.

В силу фиксированного порядка слов в английском языке BLEU завышает оценки машинного перевода в случаях, когда лексические единицы совпадают, несмотря на возможные синтаксические нарушения в предложении.

Метрика BLEU также снижает общую оценку за чрезмерную краткость (Brevity Penalty). Это работает следующим образом: вводится коэффициент, который снижает итоговый балл, если длина текста-кандидата не соответствует эталонному тексту. Штраф вычисляется по следующей формуле: $К = e^{(1 - a/b)}$, где e (экспонента) ≈ 2,718, a – количество слов в эталонном тексте, b – количество слов в тексте кандидате [Митренина, Мухамбеткалиева 2021: 82] пример:

Эталон: *The cat sits on mat* (5 униграмм).

Кандидат: *Cat on mat* (3 униграммы). $e^{(1 - 5/3)} \approx 0{,}51$. Полученный результат умножаем на общий коэффициент.

Интерпретация коэффициент оценки автоматической метрики BLEU:

0–0,19 — МП перевод не справился с задачей, суть текста уловить тяжело, такой текст сложно постредактировать;

0,2–0,39 — суть текста понятна, но есть значительное количество несовпадений, такой результат подлежит постредактированию;

0,4–0,59 — суть текста понятна, перевод также подлежит постредактированию;

≥0,6 — нехарактерный коэффициент для МП, такой показатель больше характерен для перевода, выполненного человеком.

Теперь рассмотрим, как BLEU оценила тексты научно-популярной литературы, переведённые с английского языка на русский и французский.

Таблица 1. Сравнение качества перевода научно-популярного текста (2403 символа) по метрике BLEU для разных систем машинного перевода

| Английский → Французский | | Английский → Русский | |
|---|---|---|---|
| Яндекс Переводчик | 0.45 | Яндекс Переводчик | 0.11 |
| Google Translate | 0.39 | Google Translate | 0.14 |
| DeepL | 0.43 | DeepL | 0.13 |
| DeepSeek | 0.30 | DeepSeek | 0.12 |

Результаты оценки демонстрируют значительно более высокую эффективность современных систем машинного перевода при обработке языковой пары «английский – французский». Данное явление обусловлено, на наш взгляд, следующими ключевыми факторами:

1) схожесть грамматических структур предложений, которая минимизирует смысловое расхождение при переводе;
2) высокий процент общей лексики, который снижает вероятность семантических искажений.

Таким образом, применение метрики позволило не только подтвердить различия в качестве перевода для разных языковых пар, но и установить закономерность: СМП демонстрируют более высокую точность при работе с языками, обладающими структурным и лексическим сходством.

Таблица 2. Сравнение качества перевода научно-популярного текста (1500 символов) по метрике BLEU для разных систем машинного перевода

| Английский → Французский | | Английский → Русский | |
|---|---|---|---|
| Яндекс Переводчик | 0.45 | Яндекс Переводчик | 0.14 |
| Google Translate | 0.34 | Google Translate | 0.15 |
| DeepL | 0.48 | DeepL | 0.19 |
| DeepSeek | 0.31 | DeepSeek | 0.13 |

Результаты сравнительного анализа наглядно демонстрируют различия в эффективности современных систем машинного перевода при обработке научно-популярного текста. Общая оценка метрики для всех четырех протестированных систем оказалась сопоставимой, что свидетельствует о достаточно близком уровне их эффективности. Однако можно отметить небольшое превосходство таких систем, как «DeepL» и «Яндекс.Переводчик», по сравнению с «Google Translate» и «DeepSeek» при переводе на французский язык. Особый интерес представляет случай перевода с английского на русский язык, где «Яндекс.Переводчик» продемонстрировал существенно более низкие показатели.

Подводя итоги, метрика BLEU является эффективным инструментом для предварительной оценки качества машинного перевода, позволяя сравнивать

эффективность систем для разных языковых пар и типов текстов. Однако важно помнить, что она оценивает лишь формальные соответствия, не заменяя полноценной экспертной проверки. Наилучшие результаты даёт сочетание автоматической оценки и человеческого анализа.

ЛИТЕРАТУРА

Ахрамеева К. А., Герлинг Е. Ю., Мицковский Д. Ю., Прудников С. В. Использование метрики BLEU для оценки естественности текста лингвистических стегосистем // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». 2020. №2. С. 73–80.

Митренина, О. В., Мухамбеткалиева, А. Г. Как и какой перевод (не) оценивают компьютеры // Journal of Applied Linguistics and Lexicography. 2021. Vol. 3. No. 2. Pp. 77–84.

Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). 2001. Pp. 311–318.

Satanjeev B., Alon L. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments // Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (University of Michigan, 29 June 2005). Pp. 65-72. URL: https://aclanthology.org/W05-0909.pdf