УДК 81`33, 81`25, 81`32 https://doi.org/10.33910/2687-0215-2021-3-2-77-84

Как и какой перевод (не) оценивают компьютеры

О. В. Митренина ✉[1], А. Г. Мухамбеткалиева [1]

[1] Санкт-Петербургский государственный университет, 199034, Россия, г. Санкт-Петербург, Университетская наб., д. 7–9

Сведения об авторах
Ольга Владимировна Митренина, SPIN-код: 8205-0375, ResearcherID: K-2876-2013, Scopus AuthorID: 56558829300, ORCID: 0000-0002-1750-5633, e-mail: o.mitrenina@spbu.ru

Айслу Гиляжевна Мухамбеткалиева, SPIN-код: 1403-8911, ResearcherID: GPC-5204-2022, ORCID: 0000-0002-9164-0104, e-mail: mukhambetkalieva99@gmail.com

Для цитирования: Митренина, О. В., Мухамбеткалиева, А. Г. (2021) Как и какой перевод (не) оценивают компьютеры. *Journal of Applied Linguistics and Lexicography*, т. 3, № 2, с. 77–84. https://doi.org/10.33910/2687-0215-2021-3-2-77-84

Получена 26 августа 2021; принята 7 ноября 2021.

Финансирование: Исследование не имело финансовой поддержки.

Аннотация: В статье рассматриваются современные метрики оценки качества перевода, которые используются при создании и настройке компьютерных переводчиков, при соревнованиях по машинному переводу, а также при оценке работы некоторых других систем обработки естественного языка. Описываются критерии оценки качества перевода и основные методы экспертной оценки. Рассматриваются принципы работы автоматических метрик (BLEU, TER, METEOR, BERTScore, COMET и др.), их особенности, преимущества и недостатки. Авторы подчеркивают важность появления метрик BERTScore и COMET, а также объясняют популярность некоторых традиционных метрик (например, BLEU). Современные метрики оценки качества перевода дают искаженные результаты в тех случаях, когда текст содержит много выражений с непрямыми значениями: поэтические тропы, метафоры, метонимия, юмор, загадки. Общение с помощью непрямых значений предполагает человеческую способность мыслить противоречиями, они являются источником инсайта, с помощью которого Дональд Дэвидсон описывал действие метафоры, но эта область пока еще плохо поддается компьютерной обработке. Именно поэтому оценка профессиональных переводов художественных текстов с помощью метрик показывает такие низкие результаты. Дальнейшее развитие метрик должно использовать компьютерную обработку противоречий, возможно, с помощью неконсистентных логик: паракомплектной, параконсистентной и диалетической.

Ключевые слова: машинный перевод, метрики оценки машинного перевода, BLEU, n-граммные метрики, неконсистентные логики, непрямые значения, юмор, загадки, поэтические тропы, метафора, метонимия

Translations that computers (do not) evaluate and how they do it

O. V. Mitrenina ✉[1], A. G. Mukhambetkalieva [1]

[1] Saint Petersburg State University, 7–9 Universitetskaya Emb., Saint Petersburg 199034, Russia

Authors

Olga V. Mitrenina, SPIN-код: 8205-0375, ResearcherID: K-2876-2013, Scopus AuthorID: 56558829300, ORCID: 0000-0002-1750-5633, e-mail: o.mitrenina@spbu.ru

Aislu G. Mukhambetkalieva, SPIN-код: 1403-8911, ResearcherID: GPC-5204-2022, ORCID: 0000-0002-9164-0104, e-mail: mukhambetkalieva99@gmail.com

For citation: Mitrenina, O. **V.**, Mukhambetkalieva, A. G. (2021) Translations that computers (do not) evaluate and how they do it. *Journal of Applied Linguistics and Lexicography*, vol. 3, no. 2, pp. 77–84. https://doi.org/10.33910/2687-0215-2021-3-2-77-84

Received 26 August 2021; accepted 7 November 2021.

Funding: The study did not receive any external funding.

Abstract. The article discusses modern metrics for evaluating the quality of translation used in the development and tuning of MT systems, in machine translation competitions, and in evaluating the performance of some other NLP systems. The authors describe the criteria for evaluating the quality of translation and some methods of expert (human) evaluation. The article also reveals the mechanisms of automatic metrics (such as BLEU, TER, METEOR, BERTScore, COMET), their features, advantages and disadvantages. The authors emphasize the importance of BERTScore and COMET metrics and explain the popularity of some traditional metrics (e. g., BLEU). Modern metrics for the evaluation of translation quality give distorted results when the text contains numerous expressions with indirect meanings: poetic tropes, metaphors, metonymy, humor, or riddles. Communication with indirect meanings is linked with a human ability to think in contradictions. They are a source of insight and were used by Donald Davidson to describe the mechanism of a metaphor. However, communication with indirect meanings is still difficult to computerize. That is why the metric-based evaluation of professional literary translations shows poor results. Further development of metrics should use computer processing of contradictions, possibly with the help of inconsistent logics: paracomplete, paraconsistent and dialethic.

Keywords: machine translation, evaluation metrics, BLEU, n-gram metrics, inconsistent logics, indirect meanings, humour, riddles, poetical tropes, metaphor, metonymy

Человеческая и машинная оценка перевода

«Романы Курта сильно проигрывают в оригинале...» — многие помнят этот анекдот. Он не столько о Курте Воннегуте, сколько о Рите Райт-Ковалевой, его переводчице на русский язык. Самый известный ее перевод — роман Сэлинджера «The Catcher in the rye» — «Над пропастью во ржи». Спустя сорок лет большим тиражом вышел второй перевод этого же романа, сделанный Максимом Немцовым. Он был гораздо ближе к оригиналу и без купюр советской цензуры. Название тоже стало ближе к оригиналу — «Ловец на хлебном поле». Этот перевод вызвал бурные споры и гневные отзывы, а литературовед и переводчик Виктор Топоров назвал его публикацию «актом литературного вандализма». Впрочем, с этим согласились не все, потому что в литературной критике, как и в литературе, существуют одновременно разные и даже плохо переносящие друг друга школы.

О вкусах можно спорить, как это делают литературные критики, или не спорить, как предпочитают вести себя люди, ценящие свой покой, но для машинного перевода вообще не подходят оценки по вкусу. Нужно было научиться оценивать качество переводов формализовано, то есть так, чтобы перевод, сделанный одной машиной, смогла оценить

другая машина. Машины думают числами, и поэтому нужно было научиться оценивать качество машинного перевода в числах.

Но какие числа можно привязать к переводу, чтобы с помощью них определять, какой перевод лучше? Эту задачу пришлось решать в 1966 году американской государственной комиссии ALPAC (Automatic Language Processing Advisory Committee — Консультативная комиссия по автоматической обработке языка). Она должна была определить, есть ли у машинного перевода перспективы и надо ли вкладывать государственные деньги в его развитие.

Чтобы оценить качество машинных переводов, ALPAC привлекла экспертов-людей, так как автоматических систем оценки в те годы не было. Самой популярной парой для перевода были английский и русский языки, на них и проводилась проверка.

ALPAC выделила два параметра для оценки перевода отдельных предложений (Pierce, Carroll 1966, 67–70):

Понятность (intelligibility) — от 1 до 9. Этот параметр отвечал одновременно за соответствие перевода нормам языка и за его общую понятность. Оценка 9 ставилась в тех случаях, когда переведенное предложение абсолютно понятно и не содержит грамматических или стилистических ошибок. Оценка 1 означала, что смысл предложения понять невозможно. По этому параметру оценки ставились без учета оригинала, только на основе перевода.

Точность (fidelity) — от 1 до 9. Этот параметр вычислялся косвенным образом: информанту давали перевод предложения, сделанный машиной, а потом показывали предложение на языке оригинала и просили оценить, насколько оно «информативнее» по сравнению с переводом. «Очень информативный» оригинал означал низкую точность перевода.

Позже переводы стали оцениваться по двум похожим параметрам: **полнота** (adequacy) и **гладкость** (fluency). Полнота отвечает за точность перевода, а гладкость — за его соответствие нормам языка.

По результатам проверки ALPAC постановила, что перспектив у машинного перевода нет, поэтому деньги на его развитие тратить не нужно. Это привело к сокращению исследований в этой области не только в США, но и по всему миру, включая СССР.

Однако системы машинного перевода продолжали развиваться, и их разработчикам требовались инструменты дешевой и быстрой оценки качества. Оценка должна быть автоматической, поскольку невозможно после каждого изменения системы сажать группу переводчиков-людей и заставлять их заново оценивать сотни предложений.

Быстрая автоматическая оценка машинного перевода предполагает сравнение его с эталонным переводом. Для этого необходимы два элемента:

1) корпус эталонных переводов;

2) метрика (формула) для определения, насколько далеко машинный перевод отстоит от эталонного.

Система переводит набор предложений, после чего полученный результат сравнивается с эталоном. Результаты сравнения описываются числовыми значениями. Чем ближе результат к эталону, тем лучше машинный перевод.

Собрать эталонный корпус перевода оказалось довольно простой задачей. Гораздо сложнее оказалось определиться с метриками. Как формально определить, что такое близость предложений?

Метрика BLEU

Одной из первых метрик автоматической оценки качества машинного перевода была метрика BLEU (англ. Bilingual Evaluation Understudy), разработанная сотрудниками компании IBM в 2002 году (Papineni et al. 2002). Она является одной из самых простых в использовании, а также самой популярной (несмотря на все ее недостатки).

Идея метрики проста: машинный перевод (кандидат) сравнивается с эталонным — чем больше слов совпадает, тем выше результат.

Перевод-кандидат и эталон делятся на цепочки из *n* слов (далее — n-граммы), и рассчитывается отношение количества совпадений n-грамм в кандидате и эталоне и общего количества n-грамм в **кандидате**, например:

Предложение на языке оригинала: Машинный перевод становится лучше!

Кандидат*: **Machine translation** gets **better!***

Эталон*: **Machine translation** is getting **better!***

$$\text{Точность (по униграммам)} = \frac{\text{кол-во совпадений n-грамм в кандидате и эталоне}}{\text{всего } n\text{-грамм в кандидате}} = \frac{3}{4}$$

Неплохо, правда? Но что, если машина решила сгенерировать последовательность из повторяющихся слов?

Кандидат*: **Machine machine machine machine!***

Эталон*: **Machine** translation is getting better!*

$$\text{Точность (по униграммам)} = \frac{4}{4}$$

Результат показывает, что у нас получился идеальный перевод, но с человеческой точки зрения он совершенно ужасен!

Чтобы избежать подобного, создатели BLEU решили использовать модифицированную точность, которая как бы «обрезает» количество совпадений в кандидате и эталоне, основываясь на максимальном количестве n-грамм в эталоне (Papineni et al. 2002). В таком случае мы получаем:

Кандидат: ***Machine machine machine machine!***

Эталон*: **Machine** translation is getting better! (Слово «machine»* встречается всего 1 раз)

$$\text{Точность (по униграммам)} = \frac{1}{4}$$

Теперь результат ниже и гораздо справедливее!

Еще одной проблемой является то, что BLEU не учитывает порядок слов при сравнении переводов.

Рассмотрим следующий пример:

Предложение на языке оригинала: Машинный перевод становится лучше!

Кандидат*: **Better machine translation** gets!*

Эталон*: **Machine translation** is getting **better!***

$$\text{Точность (по униграммам)} = \frac{3}{4}$$

Несмотря на то, что у кандидата и эталона много совпадений, машинный перевод все же кажется странным (возможно, не для мастера Йоды). Чтобы решить эту проблему, ввели подсчет точности для n-грамм от 1 до 4. Показатели точности затем усредняются. Посчитаем точность для n-грамм от 1 до 4 для предыдущего перевода:

Кандидат*: Better **machine translation** gets!*

Эталон*: **Machine translation** is getting better!*

Всего в кандидате 3 биграммы: *better machine, machine translation, translation gets*, а совпадений по биграммам — одно (***machine translation***). Тогда точность по биграммам = 1/3.

Если мы посчитаем точность для цепочек из 3 и 4 слов, точность составит 0, так как в кандидате нет совпадений по 3- и 4-граммам. В таком случае усредненное значение точности по n-граммам от 1 до 4 будет очень малым.

Также для случаев, когда машинный перевод намного длиннее или короче эталонного перевода, предусмотрен штраф. Так, если перевод-кандидат длиннее эталонного, вводится штраф, равный единице. Для слишком коротких переводов штраф намного меньше:

$$e^{(1-\frac{r}{c})}$$

где r — сумма длин предложений, совпадающих по длине с одним из эталонов, c — общая длина корпуса-кандидата (Papineni et al. 2002).

Таким образом, несмотря на всю свою популярность, метрика BLEU имеет множество недостатков: она не учитывает лексическое сходство переводов, высоко оценивает дословный перевод, подходит для текстов большого объема, ее результаты трудно интерпретировать и т. д.

Другие метрики

Вслед за BLEU появилось множество других метрик. Например, метрика TER (Translation Edit Rate) измеряет «редакционное расстояние», т. е. подсчитывает минимальное количество правок, необходимых для изменения машинного перевода, чтобы он точно соответствовал одному из эталонных переводов (Snover et al. 2006). Эти правки включают вставку, удаление и замену отдельных слов, а также сдвиг последовательностей слов.

TER подсчитывается следующим образом: сумма количества правок делится на общее количество слов в кандидате. Соответственно, чем ниже показатель, тем ближе машинный перевод к эталонному. Данная метрика очень полезна в задаче постредактирования машинного перевода.

Некоторые метрики позволяют подключать дополнительные функции: так, метрика METEOR использует специальный модуль для сопоставления синонимов в переводах (Banerjee, Lavie 2005). Метрика METEOR представляет собой улучшенную версию BLEU. Как мы помним, метрика BLEU суммирует точные совпадения по n-граммам. METEOR, в свою очередь, устанавливает соответствия в несколько этапов: сначала находятся точные совпадения, затем — совпадения по основам слов, и только потом — синонимы (Banerjee, Lavie 2005).

Существует еще огромное множество метрик, которые активно применяются в индустрии машинного перевода: NIST, RIBES, hLEPOR, ChrF, GLEU и многие другие. Но сейчас мы немного поговорим о тех особенных метриках, которые появились на рынке совсем недавно и успели завоевать популярность.

В наше время крупные компании осуществляют машинный перевод с помощью искусственных нейронных сетей, которые хорошо «понимают» синтаксическую структуру и семантические связи в тексте на языке оригинала и правильно отражают их в тексте перевода. Так почему же не использовать нейронные сети и для оценки качества машинного перевода?

В 2020 году компания Google представила метрику BERTScore, которая вычисляет семантическую близость между машинным и эталонным переводами (Zhang et al. 2020). BERTScore использует данные, полученные из заранее обученной нейросетевой модели-трансформера BERT (Bidirectional Encoder Representations from Transformers). Вопрос о том, как именно рассчитывается семантическая близость, заслуживает отдельной статьи.

Скажем только, что те из вас, кто задавался вопросом: «Ну и где мне пригодятся знания о косинусе?», удивятся, узнав, насколько эти знания важны!

Последняя метрика, которую хочется упомянуть, называется COMET (Crosslingual Optimized Metric for Evaluation of Translation). В 2019 году сотрудники лаборатории Unbabel AI разработали метрику COMET, которая помимо машинного и эталонного переводов принимает на вход текст на языке оригинала (Rei et al. 2020). COMET, как и BertScore, является нейросетевой моделью, и по состоянию на 2022 год появилось целое семейство метрик COMET. Среди них есть даже метрики, которые сравнивают оригинал с машинным переводом, то есть не используют эталонный перевод.

COMET и BertScore дают высокую корреляцию с человеческой оценкой, но используются не так часто, как BLEU или TER. Несмотря на все свои недостатки, такие метрики, как BLEU или TER, не требуют большой вычислительной мощности и давно используются компаниями на различных соревнованиях по машинному переводу. Однако все мы понимаем, что машинам нужно уметь переводить нечто большее, чем просто слова.

Оценка переводов художественных текстов

Лучшие художественные переводы получают крайне низкие оценки от автоматических метрик, потому что по составу слов эти переводы нередко сильно отличаются от оригиналов (Арутюнова 2018). Читатели-люди оценивают работу художественных переводчиков лучше, чем ее оценивают машины. Перевод Райт-Ковалевой «Над пропастью во ржи» считается классикой, а гораздо более близкий к оригиналу «Ловец на хлебном поле» восторгов не вызывает, хотя он гораздо «ближе» к оригиналу и получил бы более высокую оценку математической метрики. Видимо, совпадений слов для качественного художественного перевода недостаточно.

В. Н. Комиссаров определял перевод как «вид языкового посредничества, при котором содержание иноязычного текста оригинала передается на другой язык путем создания на этом языке коммуникативно равноценного текста» (Комиссаров 2002, 53). Коммуникативно равноценный перевод достаточно легко создается для текстов официально-делового стиля, в котором почти полностью отсутствуют эмоционально-экспрессивные речевые средства и индивидуализация стиля. Автоматическая оценка таких переводов традиционно показывает хорошие результаты. На другом полюсе находится язык художественной прозы и поэзии. Он содержит выразительно-изобразительные средства языка, активно использующие непрямые значения слов: стилистические фигуры, метафору, метонимию, юмор, загадки и др. Они не просто передают какое-то другое, «непрямое» значение, но и приводят к некому озарению, инсайту, с помощью которого описывал действие метафоры Дональд Дэвидсон: «Метафора заставляет нас видеть одну вещь как другую с помощью какого-нибудь прямого утверждения, которое вызывает или подсказывает инсайт» (Davidson 1984, 263).

Непрямые значения ведут к возникновению противоречий, которые могут быть описаны с помощью неконсистентной логики, т. е. логики, подразумевающей противоречия. Существует несколько ее разновидностей:

1. Паракомплектная логика (отменяет принцип исключенного третьего — допускает одновременную истинность пропозиций «все *S* суть *P*» и «никакие *S* не суть *P*»);

2. Параконсистентная логика (отменяет принцип непротиворечия — допускает, что A идентично X, B идентично тому же X, но при этом A и B не идентичны);

3. Не-алетическая логика (одновременно отменяет принцип исключенного третьего и принцип непротиворечия).

В исследовании (Лурье, Митренина 2020) показано, что метафора соответствует параконсистентной логике, метонимия — паракомплектной. Поэтический троп строится

от данности некоего объекта *a* к неожиданности некоего объекта *b*, а в загадке путь противоположный: от некоего неожиданного объекта *b* нужно перейти к данности объекта *a*. Различные нарушения правил построения двух этих элементарных непрямых значений позволяют создавать другие виды языковой игры (Лурье, Митренина 2020).

Дальнейшее развитие метрик должно использовать не только более глубокий семантический анализ, но и компьютерную обработку противоречий, возможно, с помощью неконсистентных логик: паракомплектной, параконсистентной и диалетической.

Вклад авторов

Авторы в равной мере работали над статьей, хотя описание метрик в большей степени написано А. Г. Мухамбеткалиевой, а остальная часть статьи — О. В. Митрениной.

Author Contributions

The authors made an equal contribution to the paper. Olga V. Mitrenina drafted the main body of the paper, while Aislu G. Mukhambetkalieva described the metrics.

Литература

Арутюнова, И. А. (2018) *Исследование автоматических метрик оценки перевода на материале профессиональных художественных переводов. Выпускная квалификационная работа. Уровень: магистратура.* СПб., Санкт-Петербургский государственный университет, 88 с.

Комиссаров, В. Н. (2002) *Современное переводоведение*. М.: ЭТС, 424 с.

Лурье, В. М., Митренина, О. В. (2020) Непрямые значения в естественном языке и неконсистентные логики. *Логико-философские штудии*, т. 18, № 2, с. 71–111. https://doi.org/10.52119/LPHS.2020.66.28.005

Banerjee, S., Lavie, A. (2005) METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (eds.). *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. Ann Arbor: Association for Computational Linguistics Publ., pp. 65–72.

Davidson, D. (1984) What metaphors mean. In: *Inquiries into truth and interpretation*. Oxford: Clarendon Press, pp. 245–264.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002) BLEU: A method for automatic evaluation of machine translation. In: *ACL-2002: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia: Association for Computational Linguistics Publ., pp. 311–318. https://doi.org/10.3115/1073083.1073135

Pierce, J., Carroll, J. B. (1966) *Languages and machines: Computers in translation and linguistics*. Washington: National Academy of Sciences Publ.; National Research Council Publ., 124 p.

Rei, R., Stewart, C., Farinha, A. C., Lavie, A. (2020) COMET: A neural framework for MT evaluation. In: *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*. Philadelphia: Association for Computational Linguistics Publ., pp. 2685–2702. http://doi.org/10.18653/v1/2020.emnlp-main.213

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006) A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Cambridge: The Association for Machine Translation in the Americas Publ., pp. 223–231.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. (2020) BERTScore: Evaluating text generation with BERT. In: *ICLR 2019: International Conference on Learning Representations. 6–9 May, 2019*. [Online]. Available at: https://doi.org/10.48550/arXiv.1904.09675 (accessed 23.05.2021).

References

Arutyunova, I. A. (2018) *Issledovanie avtomaticheskikh metrik otsenki perevoda na materiale professional'nykh khudozhestvennykh perevodov. Master's Thesis (Linguistics)*. Saint Petersburg, Saint Petersburg State University, 88 p. (In Russian)

Banerjee, S., Lavie, A. (2005) METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (eds.). *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. Ann Arbor: Association for Computational Linguistics Publ., pp. 65–72. (In English)

Davidson, D. (1984) What metaphors mean. In: *Inquiries into truth and interpretation*. Oxford: Clarendon Press, pp. 245–264. (In English)

Komissarov, V. N. (2002) *Sovremennoe perevodovedenie*. Moscow: EST Publ., 424 p. (In Russian)

Lourié, B. M., Mitrenina, O. V. (2020) Nepryamye znacheniya v estestvennom yazyke i nekonsistentnye logiki [Indirect meanings in natural language and inconsistent logic]. *Logiko-filosofskie studii*, vol. 18, no. 2, pp. 71–111. https://doi.org/10.52119/LPHS.2020.66.28.005 (In Russian)

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002) BLEU: A method for automatic evaluation of machine translation. In: *ACL-2002: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia: Association for Computational Linguistics Publ., pp. 311–318. https://doi.org/10.3115/1073083.1073135 (In English)

Pierce, J., Carroll, J. B. (1966) *Languages and machines: Computers in translation and linguistics*. Washington: National Academy of Sciences Publ.; National Research Council Publ., 124 p. (In English)

Rei, R., Stewart, C., Farinha, A. C., Lavie, A. (2020) COMET: A neural framework for MT evaluation. In: *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*. Philadelphia: Association for Computational Linguistics Publ., pp. 2685–2702. http://doi.org/10.18653/v1/2020.emnlp-main.213 (In English)

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006) A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Cambridge: The Association for Machine Translation in the Americas Publ., pp. 223–231. (In English)

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. (2020) BERTScore: Evaluating text generation with BERT. In: *ICLR 2019: International Conference on Learning Representations. 6–9 May, 2019*. [Online]. Available at: https://doi.org/10.48550/arXiv.1904.09675 (accessed 23.05.2021). (In English)