Проблема оценки качества дословного перевода фразеологизмов при применении автоматических метрик машинного перевода (на примере метрики COMETKiwi)

Булдаков Егор Викторович

Студент

Факультет иностранных языков и регионоведения МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: egor.buldakov@mail.ru

Научный руководитель *— Устинова Татьяна Викторовна*

Использование автоматических метрик машинного перевода, не требующих референтный перевод при оценке качества гипотетического машинного перевода, приобретают все большую актуальность, поскольку оценка таких метрик не зависит от качества референтного перевода [1]. Однако автоматические метрики такого типа могут демонстрировать предвзятость к определенным явлениям в переводе [4]. Цель данного исследования – выяснить, существует ли у метрики COMETKiwi предвзятость к дословному переводу фразеологизмов [2; 5].

Для проверки работы автоматической метрики COMETKiwi использовался датасет Idioms-InCtx-MT [3]. Этот датасет содержит набор из 1000 предложений на русском языке, содержащих фразеологизмы, и 1000 корректных переводов этих предложений на английский язык [3]. Для проведения исследования эти 1000 переводов были вручную отредактированы: корректный перевод фразеологизмов в каждом предложении был заменены на дословный перевод соответствующих фразеологизмов на русском языке. Ряд предложений, в которых корректный перевод фразеологизма на английский язык близок к дословному были исключены (например фразеологизм «и мухи не обидит» и его перевод на английский язык : wouldn't hurt a fly). Корректные и отредактированные переводы были оценены с помощью метрики COMETKiwi. С помощью набора корректных и некорректных переводов для метрики COMETKiwi была подсчитана попарная точность A:

$$A = \frac{|\text{COMETKiwi}(s, \text{cor}) > \text{COMETKiwi}(s, \text{lit})|}{|\text{all pairs}|}, \qquad (1)$$

где s — набор исходных предложений, cor — набор предложений с

корректным переводом фразеологизма, lit — набор предложений с дословным переводом фразеологизма, all pairs — все пары предложений. Результаты оценки использования метрики COMETKiwi представлены в таблице ниже:

Показатель	Значение
Средняя оценка корректных предложений	0.7117
Средняя оценка отредактированных предложений	0.7356
$\lvert$COMETKiwi(cor) > COMETKiwi(lit)$\rvert$	294
$\lvert$COMETKiwi(cor) < COMETKiwi(lit)$\rvert$	418
$\lvert$COMETKiwi(cor) > COMETKiwi(lit)$\rvert$ (%)	41.29
$\lvert$COMETKiwi(cor) < COMETKiwi(lit)$\rvert$ (%)	58.71

Таблица 1: результат оценки работы метрики

Поскольку отредактированные предложения отличаются только переводом фразеологизмов, следует получить сведения о разностях оценок предложений в парах, так как разности могут быть незначительными. Для этого следует рассчитать абсолютные разности (Δ_{abs}) между оценками. Однако, поскольку фактический диапазон оценок у COMETKiwi неясен, рассчитаем также нормализованную разность для каждой пары cor_i и lit_i:

$$\Delta_{norm} = \frac{|\text{COMETKiwi}(\text{s}_\text{i}, \text{cor}_\text{i}) - \text{COMETKiwi}(\text{s}_\text{i}, \text{lit}_\text{i})|}{R} \times 100, \quad (2)$$

где s_i — исходное предложение, R — это модуль разности между максимальной и минимальной оценками в имеющемся наборе пар. Результаты представлены в таблице ниже:

Набор данных	Δ_{abs}	Δ_{norm}	R
Все пары	0.0678	13.8564	0.4896
COMETKiwi(cor) > COMETKiwi(lit)	0.0532	12.394	0.4289
COMETKiwi(cor) < COMETKiwi(lit)	0.07815	15.964	0.4896

Таблица 2: средняя нормализованная и абсолютная разности метрики COMETKiwi

Набору пар, где метрика более высоко оценивает предложения с дословным переводом фразеологизма, соответствует более высокая

нормализованная разность, и в этом наборе пар присутствуют как наибольшая, так и наименьшая оценка для всех пар. Иными словами, для пар, где дословный перевод фразеологизма оценен выше корректного, метрика COMETKiwi в среднем устанавливает несколько большее различие между двумя оцениваемыми предложениями.

Таким образом, отсутствие различий между предложениями в паре кроме перевода фразеологизмов, большее среднее значение оценки некорректных переводов, низкая попарная точность и большее значение разности для некорректно оцененных пар свидетельствуют о предвзятости метрики COMETKiwi к дословному переводу фразеологизмов. Метрика COMETKiwi действительно в ряде случаев дает более высокую оценку предложению с дословным переводом фразеологизмов.

Литература

1. Freitag M. et al. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent // In Proceedings of the Eighth Conference on Machine Translation, 2023, P. 578–628.
2. Rei R. et al. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task // In Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, P. 634–645.
3. Stap D. et al. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities // In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024, P. 634–645.
4. Zaranis E. et al. Watching the watchers: Exposing gender disparities in machine translation quality estimation // In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), 2025, P. 25261–25284.
5. Модель COMETKiwi (Unbabel/wmt22-cometkiwi-da) на ресурсе Hugging Face: `https://huggingface.co/Unbabel/wmt22-cometkiwi-da`