УДК 81’25

Елизарова
Людмила Вячеславовна
кандидат филологических наук,
доцент кафедры перевода
Российский государственный
педагогический университет
им. А.И. Герцена
г. Санкт-Петербург, Россия

Lyudmila Elizarova
PhDin Philology,
Associate Professor
of the Translation
and Interpreting Department
Herzen State Pedagogical University
of Russia
Saint-Petersburg, Russia
lyudmilaelizarova@yandex.ru

Дмитриева
Ксения Константиновна
бакалавр
Российский государственный
педагогический университет
им. А. И. Герцена
г. Санкт-Петербург, Россия

Kseniya Dmitrieva
Bachelor’s Study
Herzen State Pedagogical University
of Russia
Saint-Petersburg, Russia
kdmitrieva2002@gmail.com

НЕЙРОННЫЕ СЕТИ VS. ТЕСТОВЫЕ МАТЕРИАЛЫ ДЛЯ ОЦЕНКИ МЕТРИК

Статья посвящена изучению особенностей нейросетевых метрик автоматической оценки качества перевода. Вследствие широкого использования программ машинного перевода растет потребность в его быстрой и адекватной оценке. Для решения этой задачи активно разрабатываются метрики автоматической оценки, сильными сторонами которых являются скорость, доступность и объективность. Качественным прорывом, определяющим в настоящее время развитие методов оценки, стал переход от традиционных метрик, использующих статистическое сравнение перевода с эталоном, к обучаемым нейросетевым метрикам. Это позволило значительно повысить гибкость и адекватность оценки метрик. Вместе с тем усложнился процесс оценки работы метрик, увеличилось количество требований к тестовым материалам. В статье рассматриваются изменения тестовых мате-

риалов на примере корпусов текстов Конференции по машинному переводу (Workshop on Machine Translation) за 2017–2022 гг. Выделяются основные факторы, влияющие на оценку работы нейросетевых метрик, указываются вариативные и инвариантные параметры текстов. В статье также приводятся практические рекомендации по отбору тестовых материалов.

К л ю ч е в ы е с л о в а: *машинный перевод; оценка качества; метрики автоматической оценки качества; нейросетевые метрики; тестовые материалы; оценка работы метрик.*

NEURAL NETWORKS VS. TEST MATERIALS FOR METRIC EVALUATION

The article is devoted to the study of the neural metrics for automatic assessment of translation quality. With the widespread use of machine translation programs, the need for its rapid and adequate evaluation is constantly growing. This leads to an increased interest in metrics, the strengths of which are speed, accessibility and objectivity. A qualitative breakthrough that currently determines the development of assessment methods has been the transition from traditional metrics using statistical comparison of translation with a reference translation to trained neural metrics. This has significantly increased the flexibility and adequacy of metric evaluation. At the same time the process of evaluating the metrics has become more complicated, and the number of requirements for test materials has increased. The article discusses changes in test materials using the example of the corpus of texts of the 2017–2022 Workshop on Machine Translation. The main factors influencing the evaluation of neural network metrics are highlighted, variable and invariant parameters of texts are indicated. The article also provides practical recommendations on the selection of test materials.

K e y w o r d s: *machine translation*; *quality assessment*; *automatic metrics; neural metrics; test materials*; *evaluation of metrics.*

С расширением сферы машинного перевода (МП) актуальным становится вопрос оценки его качества. Экспертная, или человеческая оценка, которая заключается в оценке экспертом переводов с помощью различных типологий ошибок, на данный момент считается «золотым» стандартом. Однако она занимает много времени, из-за чего не справляется с современными запросами переводческой отрасли, прежде всего, в части оценки большого объема материалов. По этой причине все чаще используются метрики автоматической оценки. Они представляют собой программы, которые сравнивают перевод с эталонным переводом, выполненным профессиональным переводчиком, или с заранее подготовленными примерами, на которых они обучались, по заданным параметрам. К сильным качествам метрик относятся скорость, доступность, а также максимальная объективность оценки, поскольку метрики не могут отойти от набора параметров. В оценке качества МП с использованием метрик можно выделить несколько подходов.

Метрики автоматической оценки развиваются параллельно с программами МП, т.к. не только тесно связаны с МП, но и часто используют те же технологии. Одним из первых подходов был статистический, в результате которого появились программы статистического машинного перевода (SMT). Подход основан на сравнении оригинала с вариантами в таблице переводов МП и выборе наиболее статистически подходящего варианта перевода. Данный принцип используется и в традиционных метриках оценки качества

(например, BLEU, chrf, NIST, ROUGE и т.д.). Метрики сравнивают перевод, предоставленный им на оценку, с эталонным переводом. Затем с помощью математических формул они выводят оценку, основанную на количестве совпадений последовательностей слов или символов (n-грамм).

Со временем стали появляться различные программы нейронного машинного перевода (NMT). Резкий подъем в их продвижении и использовании приходится на 2014–2016 гг. Суть нейросетевого подхода заключается в приближении процесса перевода к человеческому. Модель МП обучается на специально отобранных материалах языка оригинала и языка перевода, проходит этап тренировки (или тренинга) и позже в работе не просто выбирает самый статистически подходящий вариант из таблицы, а использует полученные знания о том, как пишутся тексты на обоих языках. Немного позже появляются нейросетевые метрики оценки качества. Первой нейросетевой метрикой считается ReVal, которая была предложена в 2015 году [1]. Однако бум нейросетевых метрик приходится на 2019–2020 гг. (YiSi, BLEURT, Prism, COMET). Нейросетевые метрики так же, как NMT, копируют процесс работы человека. Они проходят этапы обучения, тренировки и использования полученных знаний. В отличие от традиционных нейросетевые метрики учитывают оригинал перевода при выведении оценки. С учетом этих особенностей, нейросетевые метрики больше коррелируют с человеческой оценкой и показывают более адекватные результаты оценки МП, чем традиционные метрики.

Анализ отчетов и материалов задания «Metrics Task» Конференции по машинному переводу (WMT) с 2017 по 2022 г. [2; 3; 4; 5; 6; 7] показал, что внедрение нейронных сетей привело к необходимости изменения и более пристального подбора как учебных, так и тестовых материалов. При этом особое внимание уделяется последним. В случае, если предоставленный для оценки перевод будет полностью или частично совпадать с учебными материалами, на которых обучались нейросетевые метрики, то они могут автоматически выдавать результаты оценки, которые запомнили в процессе обучения, т.е. без анализа текста. Вследствие этого для адекватной оценки работы метрик необходимо постоянно вносить изменения в корпус используемых тестовых материалов. Так, в 2022 году на Седьмой конференции по машинному переводу (WMT22) в качестве тестовых материалов был предложен уникальный корпус текстов. Если до этого работа метрик оценивалась преимущественно на публицистических текстах (news task), то корпус 2022 года состоял из гибридных текстов, которые включают характеристики различных типов текста. Тексты были отнесены к сферам: новостного дискурса (news), электронной коммерции (e-commerce), устной разговорной речи (conversation) и к социальной сфере (social) [7]. Каждый из тестовых материалов обладал собственными уникальными чертами. Так, в текстах сферы новостного дискурса, помимо черт сугубо новостной публицистики, присутствовали особенности таких типов текста, как

• спортивный репортаж: спортивная терминология (*a simple penalty at the post*), образные выражения (*A rollercoaster first half, 13 men were on the ropes*), профессионализмы (*sin-binned*);

• научно-популярная статья: терминология (*mutations in the spike, nucleic acid-binding reactions*), аббревиатуры (*PCR test*), объяснение терминологии и сложных узкоспециальных понятий (*target DNA (mutations such as spike protein*).

В текстах сферы электронной коммерции, кроме черт технических текстов, которые присутствовали в каждом тексте корпуса, отдельно наблюдались особенности таких типов текстов, как

• рекламный текст: оценочная лексика (*has been outperforming the competition, more than any other security company*), слоганы компаний (*a protection promise only Norton can make*);

• устный деловой текст: фразы вежливости (*Please, Let me know, Ok, please do me the favour*), обращения к клиенту (*After that you will need, you chose it to be picked up by you*), упрощенная лексика и синтаксис, отсутствие специализированной терминологии (*your eReader remembers*).

Таким образом, каждый текст в корпусе представляет собой уникальный материал и существенно отличается от типовых учебных текстов, на которых обучались метрики. Гибридность и уникальность текстов позволяют адекватно оценить работу нейросетевых метрик на практике, их способность оценивать индивидуальные переводческие решения. Вследствие этого параметр гибридности представляется одним из самых важных при выборе тестовых материалов. Однако он не является единственным. При выборе тестовых материалов для оценки работы нейросетевых метрик важно учитывать следующие факторы: развитие программ МП, используемые технологии обучения, инвариантные и вариативные параметры текстов.

Качество перевода используемой программы МП непосредственно повлияет на достоверность оценки работы метрики, т. к. чаще всего в качестве тестовых материалов привлекается перевод текстов, выполненный с помощью какой-либо системы МП. Это происходит по нескольким причинам. Во-первых, машинный перевод более доступен, чем профессиональный, выполненный человеком, и разнообразие программ МП позволяет получать несколько вариантов проверяемых переводов за короткое время. Во-вторых, в то время как переводчик в силу человеческого фактора будет допускать разнообразные ошибки, ошибки программ МП будут однотипными и постоянными, т. к. программа работает по определенно заданному алгоритму и априори менее гибкая, чем профессиональный переводчик. МП больше подходит для анализа работы метрик оценки качества с определенными группами ошибок. Однако при использовании текстов МП как тестового материала необходимо учитывать некоторые особенности. Так, если в одном сегменте перевода будет сконцентрировано большое количество ошибок, то по низким общим оценкам невозможно будет определить способность метрики работать с отдельными типами ошибок, поскольку метрики выставляют

оценку всему анализируемому сегменту в целом. Для достоверного анализа работы метрики подходят сегменты перевода с минимальным количеством ошибок. Более того, необходимо учитывать материалы, на которых обучалась NMT программа. Если учебные материалы метрики и программы МП совпадут, то метрика заведомо будет оценивать данный перевод адекватнее, т. к. в ее памяти представлены варианты перевода МП с присвоенными им оценками. Но в таком случае проблематичным представляется дальнейшее развитие технологий перевода и оценки. В связи с этим работа нейросетевых метрик оценивается с использованием переводов, полученных с помощью нескольких программ МП, желательно с разными архитектурами.

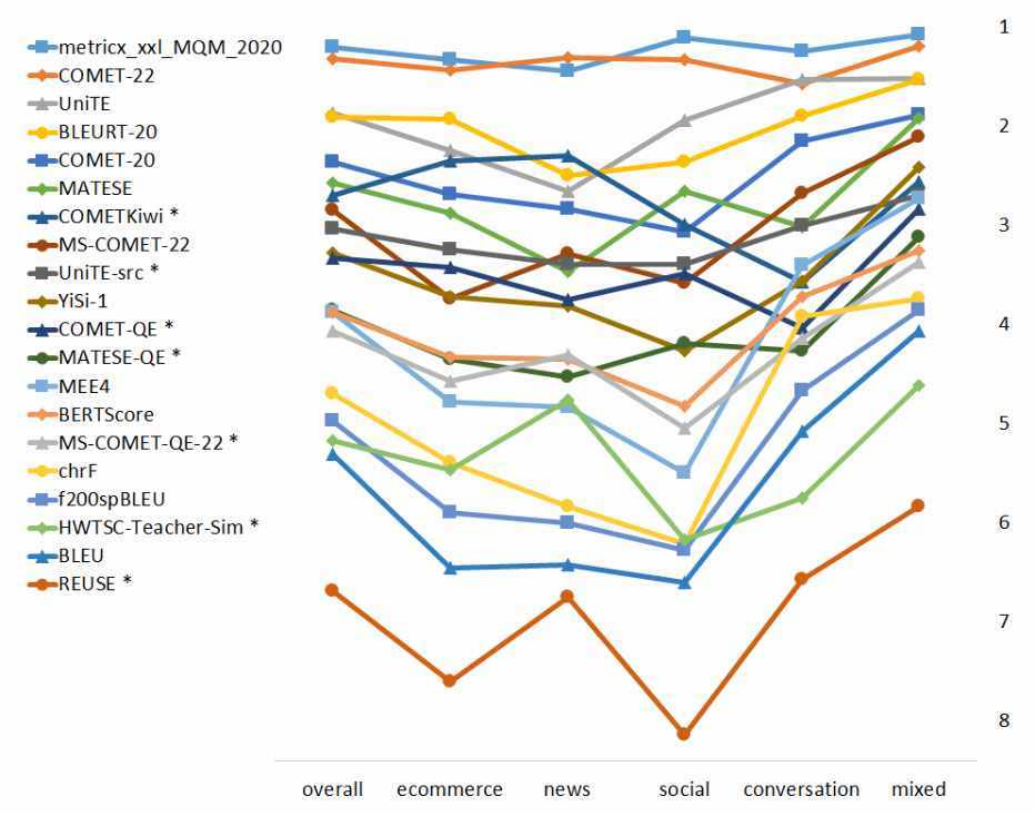
Еще одним фактором, влияющим на выбор тестового материала, являются используемые метрикой модели. На данный момент нейросетевые метрики подразделяются на референсные, которым для работы необходимо наличие эталона для сравнения, и безреференсные (reference-free metrics), которые оценивают перевод, опираясь сугубо на предобученные нейронные сети. Референсные метрики (BLEURT, Prism, UniTE) в процессе оценивания, прежде всего, опираются на предоставленный эталонный перевод, память нейросетевой модели второстепенна. Вследствие этого использование тестового материала, во многом совпадающего с эталоном при оценке их работы, не представляется результативным. Более достоверные результаты оценки можно получить, используя тестовый материал, который будет совпадать с эталоном на семантическом уровне, но не лексическом или синтаксическом. Лексическое несовпадение может достигаться за счет синонимии, использования различных приемов перевода. Синтаксическое несоответствие достигается за счет разного тема-рематического членения высказывания в таких языках, как русский и английский, изменения порядка слов, использования различных синтаксических конструкций, использования двойного отрицания в русском как способа передачи положительного смысла и т. д.

При выборе тестового материала следует также учитывать текстовые параметры. По степени влияния на достоверность оценки работы метрик их можно подразделить на инвариантные и вариативные. Инвариантные параметры являются обязательными для достижения адекватной оценки и должны присутствовать в каждом тексте из тестового корпуса. К ним относятся гибридность и разнообразие тематик.

Параметр гибридности тестовых материалов заключается в невозможности их типологизации и их отличии от типовых учебных материалов, на которых обучаются и тренируются нейросетевые метрики. В результате метрики не могут соотнести встречающиеся в гибридных текстах индивидуальные переводческие решения с типовыми вариантами перевода из корпуса учебных материалов. Благодаря этому можно оценить непосредственно работу метрик при анализе и оценке перевода, а не при простом сравнении лексических элементов в переводе и корпусе с выставлением заранее

заданных значений оценки из того же корпуса. Таким образом, параметр гибридности текстов является залогом достоверной оценки, поскольку предоставляет возможность оценить процесс работы нейросетевых моделей.

Параметр разнообразия тематик заключается в необходимости поддержания разнообразия в тематике текстов. В отличие от программ МП, которые могут создаваться и обучаться для перевода текстов отдельной тематики и работы с узкоспециальной терминологией, главная задача разработки метрик автоматической оценки заключается в повышении их гибкости и возможности учитывать как можно больше вариантов возможных переводческих решений. Однако, несмотря на то, что с внедрением нейронных сетей повысилась их гибкость, результаты работы метрик существенно разнятся в зависимости от тематики оцениваемых переводов. Ниже представлен график корреляции метрик с человеческой оценкой по данным конференции WMT22, сгруппированный по тематике переводов [7].



Тематическая корреляция с человеческой оценкой

Как видно на графике, только малый процент метрик (xxl_MQM_2020, COMET-22, COMETKiwi) показывают схожие результаты при работе с текстами разных тематик. Вследствие этого для оценки работы метрики использования материалов одной тематики будет недостаточно, данная оценка не предоставит достоверных результатов.

Наличие вариативных параметров является необязательным для каждого текста, но желательным для корпуса в целом. В результате анализа тестовых материалов “newstask” 2017–2021 гг. и “generaltask” 2022 г., а также отчетов с

соревнований работы метрик, проводимых на конференции WMT в 2017–2022 гг. [2; 3; 4; 5; 6; 7], можно выделить несколько вариативных параметров, которые в большинстве случаев связаны с единицами, вызывающими проблемы при их оценке.

1. *Проблема оценки переводов лексических единиц с существующими переводными эквивалентами.* Благодаря внедрению нейронных сетей метрики стали учитывать больше вариантов переводческих решений, более адекватно оценивать синонимию в переводе. Однако из-за невозможности учитывать в полной мере контекст метрики стали более лояльно относиться к ошибкам в переводе единиц с существующими переводными эквивалентами: терминология, устоявшиеся варианты перевода имен собственных (топонимы, названия организаций) и т. д.

2. *Проблема оценки передачи имен собственных возникают в результате отсутствия одного определенного принципа их передачи.* При переводе используются приемы транслитерации, транскрипции, калькирования, транспозиции, причем выбор приема чаще всего зависит от контекста и экстралингвистических, прагматических факторов ситуации перевода, которые метрики не могут учесть при анализе и выставлении оценки. Более того, ввиду отсутствия системы транскрипции и транслитерации вариативность возникает и в случае использования определенного приема. При отсутствии эталона, в случае с безреференсными метриками, данная проблема становится более заметной, поскольку метрики оценивают различные варианты передачи имен собственных как синонимические.

3. *Проблема оценки перевода культуроспецифичных единиц*. Проблемы с оценкой переводов данных элементов нейросетевыми метриками можно подразделить на два вида. Первый – это неспособность метрики оценить передачу самой единицы. Чаще всего это объясняется недостаточным количеством примеров употребления культуроспецифичной единицы в текстах на языке перевода в корпусе учебных материалов. Это приводит к неспособности метрики выделить и идентифицировать единицу, а также адекватно проанализировать все варианты перевода данной единицы на языке перевода. Вторая проблема связана с форматом передачи информации: адресов, дат, единиц измерения, валют и других культуроспецифичных элементов, которые необходимо адаптировать при переводе.

К указанным проблемам относятся лексические элементы, которые вызывают наибольшую трудность у метрик автоматической оценки. Наличие данных элементов в тестовых материалах позволит адекватно оценить работу метрик, а также проанализировать технологии работы с проблемными единицами и в перспективе выработать способы улучшения автоматической оценки.

В результате комплексного анализа данных факторов и их влияния на процесс подбора тестовых материалов предлагаются следующие рекомендации.

1. В качестве тестовых материалов использовать переводы МП.

2. Выбирать наиболее качественные, современные версии программ МП.

3. Переводить тестовые материалы с помощью нескольких программ МП.

4. При оценивании работы референсных метрик использовать тестовые материалы, которые семантически совпадают с эталоном, но различаются на лексическом и синтаксическом уровнях.

5. В целях получения достоверных результатов осуществлять оценку работы метрик на материале гибридных текстов из реальной переводческой практики.

6. В корпус текстов включать тексты различных тематик.

7. Использовать тексты, представляющие трудность для нейросетевых метрик в части передачи терминологии, культуроспецифичных единиц и т. д.

Развитие метрик автоматической оценки, в частности внедрение нейронных сетей, повлияло на улучшение оценки машинного перевода. Однако для дальнейшего анализа работы нейросетевых метрик, ввиду фактора их обучаемости, необходимо постоянно вносить изменения в корпус тестовых материалов. В связи с этим встает вопрос о принципах отбора тестовых материалов и факторах, влияющих на достоверность оценки работы нейросетевых метрик. В исследовании на материале тестовых материалов и отчетов Конференции по машинному переводу (WMT) 2017–2022 гг. были выделены основные факторы: развитие систем МП, нейросетевые модели метрик и непосредственно текстовые параметры.

ЛИТЕРАТУРА

1. Gupta R., Orasan C., Genabith J. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks // Association for Computational Linguistics. 2015. P. 1066–1072.

2. Bojar O., Graham Y., Kamran A. Results of the WMT17 Metrics Shared Task // Association for Computational Linguistics. 2017. P. 489–513.

3. Ma Q., Bojar O., Graham Y. Results of the WMT18 Metrics Shared Task // Association for Computational Linguistics. 2018. P. 671–688.

4. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges / Q. Ma, J. T-Z. Wei, O. Bojar, Y. Graham // Association for Computational Linguistics. 2019. P. 62–90.

5. Results of the WMT20 Metrics Shared Task / N. Mathur, J. T-Z. Wei, Q. Ma, M. Freitag, O. Bojar // Association for Computational Linguistics. 2020. P. 688–725.

6. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain / M. Freitag, R. Rei, N. Mathur, C-K Lo, C. Stewart, G. Foster, A. Lavie, O. Bojar // Association for Computational Linguistics. 2021. P. 733–774.

7. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust / M. Freitag, R. Rei, N. Mathur, C-K Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, A. F. T. Martins// Association for Computational Linguistics. 2022. P. 46–68.