В. А. Нуриев, А. Ю. Егорова, Методы оценки качества машинного перевода: современное состояние, *Информ. и её примен.*, 2021, том 15, выпуск 2, 104–111

DOI: 10.14357/19922264210215

МЕТОДЫ ОЦЕНКИ КАЧЕСТВА МАШИННОГО ПЕРЕВОДА: СОВРЕМЕННОЕ СОСТОЯНИЕ

В. А. Нуриев[1], А. Ю. Егорова[2]

Аннотация: Представлен обзор современных методов оценки качества машинного перевода (МП). В основе этих методов лежат два подхода — автоматический и экспертный. Автоматическая оценка построена на сопоставлении с референтным (профессиональным/эталонным) переводом (РП). Экспертная (с привлечением человека-эксперта) учитывает в первую очередь функциональность: качество перевода оценивается в прагматико-функциональном аспекте, т. е. принимается во внимание, насколько полученный перевод справляется со своими задачами. В первой части статьи рассматривается ряд метрик, используемых для автоматической оценки МП, отмечаются их недостатки и описываются новые направления в их разработке. Вторая часть статьи сфокусирована на экспертной оценке МП. Здесь приведены несколько основных способов такой оценки: оценивание в соответствии с критериями точности и естественности, ранжирование переводов, прямое оценивание, оценка с учетом коэффициента редактирования перевода человеком, аннотирование перевода с применением типологии ошибок.

Ключевые слова: машинный перевод; качество перевода; оценка качества машинного перевода; автоматические метрики; прямое оценивание; типология ошибок машинного перевода

DOI: 10.14357/19922264210215

1 Введение

Проблема и, следовательно, необходимость оценки качества переводного текста возникает на регулярной основе, причем не только в профессиональном сообществе, но и в жизни обычного человека. Во многом это связано с тем, что МП стал неотъемлемой частью повседневной реальности. Происходит реструктуризация рынка переводческих услуг и, в частности, МП, а бюджет индустрии постоянно наращивает объемы (см. об этом [1, с. 260–263]). Ежедневно с помощью публично доступного веб-сервиса нейронного МП Google.Translate обрабатывается около 143 млрд слов в 100 языковых парах [2]. Человек использует МП для решения задач широкого профиля: для получения информации, связанной с конкретной и требующей незамедлительных действий проблемой (перевод технического сопровождения, инструкции к лекарствам и т. д.); для покупок на зарубежных сайтах; в целях оптимизации профессиональной деятельности переводчика с помощью внедрения в его рабочий цикл этапа, предполагающего последующую редактуру автоматически сгенерированного текста. Не все указанные задачи предполагают обязательный высокий уровень качества МП. Для осуществления ряда из них достаточно общего понимания содержания даже при наличии несущественных ошибок, нарушающих правила целевого языка. Для выполнения других задач требуется высокое качество полученного автоматическим способом перевода, что указывает на необходимость постоянно оценивать динамику этого качества, изучать и совершенствовать методы его оценки. Этим обусловлено повышенное внимание научного сообщества к данной проблеме: за последние четыре года были опубликованы несколько авторитетных монографий, фокусирующихся на оценке качества МП [3–5].

Целью статьи, таким образом, ставится обзор современных тенденций в разработке методов оценки качества МП. В основе этих методов лежат два подхода — автоматический и экспертный. Автоматическая оценка построена на сопоставлении с референтным (профессиональным/эталонным) переводом (*англ.* reference translation). Экспертная (с привлечением человека-эксперта) оценка учитывает в первую очередь функциональность: качество перевода оценивается тем выше, чем успешнее он справляется со своими задачами.

2 Автоматическая оценка качества машинного перевода

Как правило, автоматическая оценка измеряет уровень соответствия МП одному или нескольким РП. Чтобы определить уровень соответствия

[1] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, nurieff.v@gmail.com

[2] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ann.shurova@gmail.com

МП и РП, применяются критерии точности (доля правильно переведенного) и полноты (доля переведенных слов, совпадающих с профессиональным переводом). Ниже представлены некоторые метрики автоматической оценки качества МП.

Метрика, к которой обращаются чаще всего, — это разработанная в IBM метрика BLEU (Bilingual Evaluation Understudy) [6]. Она вошла в золотой стандарт автоматической оценки качества МП и нередко применяется в качестве эталонной. Сопоставление МП и РП проводится путем вычисления n-граммной точности (максимальная длина n-граммного блока слов равна 4). Чтобы избежать искажения в оценке, за слишком короткий перевод назначается штраф (brevity penalty — BP), n-граммная точность при этом представляет собой «отношение последовательностей из n слов, совпадающих в МП и РП, к общему числу последовательностей из n слов в МП» [7, с. 111–112]. «Оценка... вычисляется как произведение среднего геометрического из полученных модифицированных коэффициентов и штрафного коэффициента» [8, с. 86]. Полученное значение BLEU изменяется в пределах от 0 до 1. При процентном представлении значение изменяется в промежутке от 0% до 100%. Вычисляется метрика по следующей формуле:

$$\mathrm{BLEU} = \mathrm{BP}\exp\left(\sum_{n=1}^{n} w_i \log p_i\right),$$

$$\mathrm{BP} = \min\left(1, \frac{c}{r}\right),$$

«где w_i — положительные веса для каждого используемого параметра n-грамм... n — максимальная длина n-грамм, i — длина блока в пределах n-граммы, p_i — модифицированная точность n-грамм, c — длина полученного машинного перевода, r — длина наилучшего совпадающего эталонного текста» [8, с. 86]. Метрика BLEU использует статистические инструменты, не принимая во внимание лингвистические знания.

Другая наиболее востребованная метрика — METEOR (Metric for Evaluation of Translation with Explicit Ordering) — предусматривает интеграцию языковых знаний. Так, наряду с n-граммными совпадениями в МП и РП она учитывает изменения в словоформах, синонимические ряды и т. д. [9]. Поэтому для обеспечения ее функционирования необходимо привлечение баз данных, содержащих лингвистическую информацию, нужна морфологическая разметка и вычислительно затратное пословное выравнивание. Иначе говоря, требуется сложная и тонкая настройка, куда вовлечено гораздо больше параметров, чем в BLEU.

Еще одной получившей широкое распространение метрикой автоматической оценки качества МП стала TER (Translation Error Rate). Она исходит из расчета исправлений/трансформаций, необходимых для приведения МП к эталонному образцу, и вычисляется по следующей формуле:

$$\mathrm{TER} = \frac{\text{Число редактирований}}{\text{Средняя длина эталонных переводов}}.$$

При этом пунктуационные знаки принимаются за отдельные слова, а трансформациями считаются не только удаление, вставка и замена, но и перестановка — в отличие, например, от метрики WER (Word Error Rate), которая эту последнюю трансформацию не учитывает (подробнее о TER и WER см. в [8, 10]).

Наряду с рассмотренными метриками автоматической оценки качества МП также имеются: PER (Position-Independent Word Error Rate), chrF (Character F-measure), NIST (название образовано от US National Institute of Standards and Technology) и др. (о них см. [7, 8, 11, 12]).

Целесообразность использования метрик автоматической оценки качества МП постоянно ставится под вопрос. Действительно, может ли метрика с простейшим алгоритмом вычисления типа BLEU (как и другие метрики) адекватно отражать отличия между МП и РП? Основные критические замечания, высказываемые в этой связи, заключаются в следующем.

1. При вычислении не принимается во внимание, что слова несут на себе разную функциональную нагрузку и имеют неодинаковую релевантность для формирования предложения.
2. Сравнение РП и МП носит локальный характер и проводится на уровне n-граммного соответствия, при этом упускается из виду грамматическая связность в рамках всего предложения, что искажает результаты в пользу систем МП, которые лучше переводят отдельные словарные блоки, но не всегда способны грамматически правильно оформить целое предложение.
3. Вычисляемые значения не информативны: неизвестно, как интерпретировать значение BLEU, равное, например, 30,7%, так как при вычислении задействовано множество факторов — число РП, языковая пара, терминологическое наполнение текста, схема токенизации, используемая для вычленения слов в РП и МП.
4. Ненадежность алгоритма оценки. Так, недавние эксперименты показали, что BLEU оценивает выполненные человеком переводы на том же уровне, что и машинные, хотя последние имеют гораздо худшее качество. В ходе этих

экспериментов с помощью BLEU выполнялось сравнение между несколькими РП, а также между РП и МП.

Эти недостатки необходимо учитывать в разработке новых метрик автоматической оценки МП, как необходимо учитывать и то, что в идеальном случае оценка, получаемая с помощью такой метрики, должна демонстрировать явную корреляцию с оценкой человека-эксперта. Обычно эту корреляцию рассчитывают с помощью коэффициента корреляции Пирсона [10, с. 61]. Его значение варьируется от 0 до 1, и чем оно выше в указанном диапазоне, тем лучше метрика.

Автоматические метрики получили всеобщее признание в качестве эффективного способа для оценки продуктивности систем статистического МП, однако они не совсем приспособлены, чтобы сравнивать производительность систем МП разного типа между собой, и в этом отношении разработки средств автоматической оценки МП пока не достигли сколь-нибудь значимых результатов. Вместе с тем такие разработки интенсивно ведутся, и автоматические метрики постоянно совершенствуются.

Так, все большее распространение получает подход, учитывающий при сопоставлении МП и РП морфологические, синтаксические и семантические параметры. Это, например, MEANT, где сопоставляются синтаксические древовидные структуры и принимаются во внимание такие свойства, как семантические роли [13]. Или RIBES (Rank-based Intuitive Bilingual Evaluation Score) — метрика, специально разработанная для языковых пар типа японский–английский, где коренным образом различается синтаксическое устройство.

Имеются попытки применять машинное обучение — обучать метрики на данных, полученных по результатам оценки человеком-экспертом (см., например, BEER (BEtter Evaluation as Ranking) [14, 15] или BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) — одну из самых новых метрик, которая использует нейросетевую языковую модель BEURT и обучается на рейтинговых данных [16]).

Наряду с увеличением степени корреляции между оценкой человека-эксперта и автоматической оценкой МП разработчики также стремятся обеспечить большую информативность метрик (см. выше замечание о непрозрачности значения BLEU) и снижение вычислительной трудоемкости.

Важным сейчас становится создание информационных ресурсов, содержащих лингвистические знания разной направленности, предназначенные для обучения современных автоматизированных метрик. Примером таких ресурсов служат надкорпусные базы данных, разрабатываемые в отделе 54 ФИЦ ИУ РАН [17].

Подробнее о новейших разработках в области автоматизированных средств оценки качества МП см. [10, с. 59–64].

3 Экспертная оценка качества машинного перевода

Говоря об экспертной оценке качества МП с привлечением специалистов (лингвистов, переводчиков), можно выделить несколько основных способов: оценивание в соответствии с критериями точности и естественности, ранжирование переводов, прямое оценивание, оценка с учетом коэффициента редактирования перевода человеком, аннотирование перевода с применением типологии ошибок.

Понятие «правильности» перевода недоопределено и, следовательно, плохо применимо. Вот почему для оценки качества МП руководствуются критериями[1] точности (adequacy) и естественности (fluency), используя в опросе экспертов 5-балльную шкалу Ликерта [19]. Такой подход имеет свои недостатки: эксперты не всегда последовательны в своем выборе из-за неоднозначности определений в шкале оценки, к тому же одни специалисты более снисходительны при назначении оценок, чем другие.

Чтобы избежать этих трудностей, при оценке двух и более систем МП применяется ранжирование переводов относительно друг друга. Для измерения меры согласия между экспертами используют коэффициенты каппа Коэна [10, с. 48], каппа Флейса [20].

Так, в работе [21] ранжирование проводилось для оценки качества переводов, реализованных посредством системы статистического фразового МП (СФМП) и системы нейронного МП (НМП). В ходе эксперимента каждому эксперту были представлены триплеты, состоящие из предложения на исходном языке и двух его переводов (полученных с помощью СФМП и НМП). Экспертам предлагалось оценить триплет и приписать его к одному из трех классов, показывающих соотношение качества сравниваемых переводов:

$$\text{СФМП} = \text{НМП}\,;\ \text{СФМП} < \text{НМП};\ \text{СФМП} > \text{НМП}\,.$$

Полученные экспертные оценки были сопоставлены с результатами автоматических метрик (BLEU, TER, Character F-measure).

[1] Другие возможные критерии оценки описаны в [18].

В эксперименте, описанном в [22], помимо ранжирования еще задействованы постредактирование переводов, экспертная аннотация ошибок в МП, а также оценка точности/естественности. Подобно [21], для установления корреляции между экспертной и автоматической оценкой используются автоматические метрики.

Одной из последних разработок в области оценки качества МП является прямое оценивание (direct assessment) [10, с. 49–50]. Оно предполагает оценку одного предложения единовременно (в отличие от ранжирования переводов) с применением 100-балльной шкалы, которая имеет вид немаркированной прямой с бегунком. Для экспертов характерны неодинаковые ожидания в отношении качества МП: одни склонны его оценивать выше, а другие, наоборот, ниже, что может объясняться имеющимися предубеждениями о низком качестве МП. Кроме того, разными экспертами 5-балльная шкала используется неравномерно — некоторые никогда не ставят самый низкий и самый высокий баллы. 100-балльная шкала представляет собой более гибкий оценочный инструмент. Она дает возможность измерить ожидания в отношении качества МП у каждого эксперта с помощью среднего балла всех его оценок, выявляя задействованный интервал шкалы, который отражается в дисперсии оценок. Оценки разных экспертов нормируются согласно формулам в [10, с. 49–50]. Переводы, поступающие эксперту для обработки, генерируются в разных системах МП и выбираются случайным образом. После нормирования оценок, полученных от каждого из экспертов, вычисляется средний балл для переводов отдельно взятой системы МП.

Прямое оценивание было использовано в ходе краудсорсинговой кампании по оценке качества МП, организованной ACL (Association for Computational Linguistics) в 2018 г. в рамках Конференции по компьютерной лингвистике (Workshop on Machine Translation, WMT).

Оценивать качество перевода можно и с точки зрения усилий по его постредактированию. Так, при оценке МП с учетом HTER[1] (Human Translation Edit Rate — коэффициент редактирования перевода человеком) [10, с. 51–52] эксперты получают подборку переводов, выполненных разными системами МП, которые им предлагается отредактировать. Затем для каждой системы МП проводится сопоставление перевода с его отредактированной версией и подсчитывается число изменений, сделанных экспертом.

Качество МП может оцениваться и в процессе аннотирования перевода с применением типологии ошибок. Обзор классификаций представлен в работе [23].

Одной из наиболее известных является типология DQF/MQM (Dynamic Quality Framework — динамическая модель оценки качества; Multidimensional Quality Metrics — многомерные метрики качества), разработанная в TAUS[2] и DFKI[3] в 2014 г. [24]. Типология имеет 4 уровня: наиболее специфицированные типы ошибок относятся к четвертому уровню; при этом при оценке перевода можно выбирать степень спецификации, т. е. использовать от одного до четырех уровней в зависимости от задачи. Также в типологии учитываются четыре степени критичности ошибок. Подробнее о MQM-метриках см. в работе [7].

Типология DQF/MQM получила широкое распространение. Так, в [25] она применяется для проведения количественного анализа работы разных систем МП. При этом классификация претерпевает ряд изменений, обусловленных необходимостью учитывать особенности славянских языков (в данном случае хорватского).

Следует отметить, что типологии ошибок могут быть специфицированы в зависимости от цели исследования. Так, по мнению авторов статьи [26], категория «Терминология» в классификации MQM не отражает нюансы, которые могут возникать при ошибочном переводе терминов. В работе предпринята попытка провести анализ ошибок в переводе терминов, уточнить их классификацию и сопоставить на этой основе работу систем СФМП и НМП.

Представленная в [26] типология ошибок включает в себя 5 классов:

(1) «Ошибка в словопорядке» (Reorder error);

(2) «Ошибка в формообразовании» (Inflectional error);

(3) «Ошибка в части термина» (Partial error);

(4) «Лексическая ошибка» (Incorrect lexical selection);

(5) «Пропуск термина» (Term drop).

Оставшиеся виды ошибок образуют 6-й класс, который подразделяется на три подкласса:

– «Копирование исходного термина» (Source term copied);

[1] Аббревиатура совпадает с названием автоматической метрики HTER (Human-targeted Translation Error Rate) (подробнее см. [18, с. 25]).

[2] Translation Automation User Society — Пользовательское сообщество по автоматизации перевода.

[3] Deutsches Forschungszentrum für Künstliche Intelligenz — Немецкий центр исследований искусственного интеллекта.

– «Ошибка, вызванная затруднением при снятии многозначности слова на целевом языке» (Disambiguation issue in target);
– «Другие ошибки» (Other error).

Переводы исходных терминов, в которых не было допущено ошибок, объединены в отдельный класс «Правильного перевода» (Correct translation). В нем авторы исследования выделяют еще 7 подклассов, которые демонстрируют разнообразие моделей перевода и отображают степень соответствия переводного эквивалента исходному термину.

Еще одним примером типологии ошибок может послужить классификация, представленная в [27]. Она подробно описана в работе [28]. Эта типология имеет 5 укрупненных классов ошибок, которые, в свою очередь, делятся на подклассы.

В работе [29] проводится количественный анализ ошибок мультимодальных систем НМП, способных обрабатывать изображения. За основу взята классификация ошибок [27] с некоторыми уточнениями. Изменения в ней обусловлены интересом авторов исследования к тому, как мультимодальные системы НМП переводят «визуальные» термины (visual terms) — термины, обозначающие понятия, прямое соответствие которым можно найти на предъявляемом изображении, причем задействованы только укрупненные классы типологии ошибок [27] без уточнения их дальнейшего иерархического устройства.

С учетом всех преобразований модифицированная классификация [29] включает в себя следующие классы ошибок:

(1) «Пропущенные слова» (Missing words);
(2) «Неправильные слова» (Incorrect words), куда входят подклассы
   – «Неправильный перевод» (Mistranslation);
   – «Неправильная форма, лишние слова, стилистическая ошибка» (Incorrect form, extra words or style);
(3) «Другие ошибки» (Other), куда входят подклассы
   – «Словопорядок» (Word order);
   – «Неизвестные слова» (Unknown words);
   – «Пунктуационная ошибка» (Punctuation).

Также был добавлен новый, 4-й класс, получивший название «Визуальная категория» (Visual category). В него входят 4 подкласса:

– «Правильный перевод» (Correct);
– «Неправильный перевод» (Mistranslation);
– «Неправильный, но интересный перевод» (Incorrect but interesting);
– «Новый термин» (Novel).

Разнообразие способов экспертной оценки МП свидетельствует о неослабевающем интересе профессионального сообщества к этой области, а также говорит о том, что даже с учетом существенно меньшей стоимости и большей быстроты автоматической оценке не доверяют полностью и стремятся проверить ее с помощью мнения компетентного человека.

4 Заключение

В статье представлен обзор современных подходов к оценке качества МП. Выделены два основных направления: автоматизированная оценка и оценивание с привлечением человека-эксперта. С изменением парадигмы МП и внедрением нейросетей в архитектуру автоматических переводчиков изменяются и разработки в области оценки качества МП. Это затрагивает в первую очередь автоматические метрики, используемые для оценивания переводов: для обеспечения их работы пытаются применять машинное обучение. В качестве тренировочных привлекаются данные, полученные по результатам оценки человеком. Нововведения имеются и в области экспертной оценки качества МП. Одна из последних разработок здесь — прямое оценивание. Востребованным остается аннотирование МП с применением типологии ошибок. Оно стало одним из самых продуктивных способов оценивания, поскольку позволяет гибко типологизировать ошибки в соответствии с целым рядом параметров, которые легко варьировать в зависимости от конкретных характеристик текста, поступающего на вход системы МП.

Литература

1. *Larsonneur C.* Neural machine translation: From commodity to commons? // When translation goes digital: Case studies and critical reflections / Eds. R. Desjardins, C. Larsonneur, Ph. Lacour. — Cham, Switzerland: Palgrave Macmillan, 2021. P. 257–280.
2. *Davenport C.* Google Translate processes 143 billion words every day // Android Police, 2018. https://www.androidpolice.com/2018/10/09/google-translate-processes-143-billion-words-every-day.
3. Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Machine translation: Technologies and applications ser. — Cham, Switzerland: Springer International Publishing, 2018. Vol. 1. 292 p.

4. *Specia L., Scarton C., Paetzold G. H.* Quality estimation for machine translation. — Synthesis lectures on human language technologies ser. — London: Morgan & Claypool, 2018. 162 p.
5. *Bittner H.* Evaluating the evaluator: A novel perspective on translation quality assessment. — New York, NY, USA: Routledge, 2020. 282 p.
6. *Papineni K., Roukos S., Ward T., Zhu W. J.* BLEU: A method for automatic evaluation of machine translation // 40th Annual Meeting on Association for Computational Linguistics Proceedings. — Philadelphia, PA, USA: Association for Computational Linguistics, 2002. P. 311–318.
7. *Рычихин А. К.* О методах оценки качества машинного перевода // Системы и средства информатики, 2019. Т. 29. № 4. С. 106–118.
8. *Козина А. В., Черепков Е. А., Белов Ю. С.* Автоматические метрики оценки качества машинного перевода // Системный администратор, 2019. № 11. С. 84–87.
9. *Banerjee S., Lavie A.* METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics Proceedings. — Ann Arbor, MI, USA: Association of Computational Linguistics, 2005. P. 65–72.
10. *Koehn Ph.* Neural machine translation. — New York, NY, USA: Cambridge University Press, 2020. 394 p.
11. *Popović M.* chrF: Character n-gram F-score for automatic MT evaluation // 10th Workshop on Statistical Machine Translation Proceedings. — Lisboa, Portugal: Association for Computational Linguistics, 2015. P. 392–395.
12. *Popović M.* chrF deconstructed: β parameters and n-gram weights // 1st Conference on Machine Translation Proceedings. — Berlin, Germany: Association for Computational Linguistics, 2016. Vol. 2. P. 499–504.
13. *Chi-kiu Lo.* MEANT 2.0: Accurate semantic MT evaluation for any output language // Conference on Machine Translation Proceedings. — Copenhagen, Denmark: Association for Computational Linguistics, 2017. Vol. 2. P. 589–597.
14. *Stanojević M., Sima'an K.* BEER: BEtter evaluation as ranking // 9th Workshop on Statistical Machine Translation Proceedings. — Baltimore, MD, USA: Association for Computational Linguistics, 2014. P. 414–419.
15. *Stanojević M., Sima'an K.* Evaluating MT systems with BEER // Prague Bulletin Mathematical Linguistics, 2015. No. 104. P. 17–26.
16. *Sellam T., Das D., Parikh A. P.* BLEURT: Learning robust metrics for text generation // arXiv.org, 9 Apr 2020. arXiv:2004.04696 [cs.CL].
17. *Инькова О. Ю.* Надкорпусная база данных как инструмент изучения формальной вариативности коннекторов // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Международ. конф. «Диалог». — М.: РГГУ, 2018. Вып. 17(24). С. 240–253.
18. *Castilho Sh., Doherty S, Gaspari F., Moorkens J.* Approaches to human and machine translation quality assessment // Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. P. 9–38.
19. *Likert R.* A technique for the measurement of attitudes // Arch. Psychol., 1932. Vol. 140. P. 1–55
20. *Fleiss J. L.* Measuring nominal scale agreement among many raters // Psychol. Bull., 1971. Vol. 76. No. 5. P. 378–382.
21. *Shterionov D., Superbo R., Nagle P., et al.* Human versus automatic quality evaluation of NMT and PBSMT // Machine Translation, 2018. Vol. 32. P. 217–235.
22. *Castilho S., Moorkens J., Gaspari F., et al.* Evaluating MT for massive open online courses. A multifaceted comparison between PBSMT and NMT systems // Machine Translation, 2018. Vol. 32. P. 255–278.
23. *Popovic M.* Error classification and analysis for machine translation quality assessment // Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. P. 129–158.
24. *Lommel A.* Metrics for translation quality assessment: A case for standardizing error typologies // Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. P. 109–127.
25. *Klubička F., Toral A., Sánchez-Cartagena V. M.* Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian // Machine Translation, 2018. Vol. 32. P. 195–215.
26. *Haque R., Hasanuzzaman M., Way A.* Analysing terminology translation errors in statistical and neural machine translation // Machine Translation, 2020. Vol. 34. P. 149–195.
27. *Vilar D., Xu J., D'Haro L., Ney H.* Error analysis of statistical machine translation output // 5th Conference (International) on Language Resources and Evaluation Proceedings. — Genoa, Italy: European Language Resources Association, 2006. P. 697–702.
28. *Гончаров А. А., Бунтман Н. В., Нуриев В. А.* Ошибки в машинном переводе: проблемы классификации // Системы и средства информатики, 2019. Т. 29. № 3. С. 92–103.
29. *Calixto I., Liu Q.* An error analysis for image-based multimodal neural machine translation // Machine Translation, 2019. Vol. 33. P. 155–177.

Поступила в редакцию 14.04.2021

METHODS OF QUALITY ESTIMATION FOR MACHINE TRANSLATION: STATE-OF-THE-ART

V. A. Nuriev and A. Yu. Egorova

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper reviews the state-of-the-art methods of quality estimation for machine translation. These methods are grounded in two general approaches: automatic and manual. The automatic assessment builds on the data from comparison of the machine translation system output against the human-generated reference translation. The manual (human) evaluation primarily takes into account pragmatic and functional aspects: the translation quality is assessed bearing in mind how well the system output is suited to fulfill the translation tasks. The first part presents some automatic metrics for evaluation of machine translation quality. Also, it speaks about both shortcomings of such metrics and new trends in their development. The other part of the paper is focused on human evaluation of machine translation. It describes: (i) evaluation of adequacy and fluency; (ii) ranking of translations; (iii) direct assessment; (iv) computation of the human translation edit rate, and (v) translation annotation involving an error typology.

Keywords: machine translation; translation quality; evaluation of machine translation quality; automatic metrics; direct assessment; typology of machine translation errors

DOI: 10.14357/19922264210215

References

1. Larsonneur, C. 2021. Neural machine translation: From commodity to commons? *When translation goes digital: Case studies and critical reflections*. Eds. R. Desjardins, C. Larsonneur, and P. Lacour. Cham: Palgrave Macmillan. 257–280.
2. Davenport, C. 2018. Google Translate processes 143 billion words every day. *Android Police*. Available at: https://www.androidpolice.com/2018/10/09/google-translate-processes-143-billion-words-every-day/ (accessed May 5, 2021).
3. Moorkens, J., S. Castilho, F. Gaspari, and S. Doherty, eds. 2018. *Translation quality assessment: From principles to practice*. Machine translation: Technologies and applications ser. Cham: Springer International Publishing. Vol. 1. 299 p.
4. Specia, L., C. Scarton, and G. H. Paetzold. 2018. *Quality estimation for machine translation*. Synthesis lectures on human language technologies ser. London: Morgan & Claypool Publs. 162 p.
5. Bittner, H. 2020. *Evaluating the evaluator: A novel perspective on translation quality assessment*. New York, NY: Routledge. 282 p.
6. Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *40th Annual Meeting on Association for Computational Linguistics Proceedings*. Philadelphia, PA: Association for Computational Linguistics. 311–318.
7. Rychikhin, A. K. 2019. O metodakh otsenki kachestva mashinnogo perevoda [On methods of machine translation quality assessment]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(4):106–118.
8. Kozina, A. V., E. A. Cherepkov, and Yu. S. Belov. 2019. Avtomaticheskie metriki otsenki kachestva mashinnogo perevoda [Automatic metrics for machine translation evaluation]. *Sistemnyy administrator* [System Administrator] 11:84–87.
9. Banerjee, S., and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics Proceedings*. Ann Arbor, MI: Association of Computational Linguistics. 65–72.
10. Koehn, Ph. 2020. *Neural machine translation*. New York, NY: Cambridge University Press. 394 p.
11. Popović, M. 2015. chrF: Character n-gram F-score for automatic MT evaluation. *10th Workshop on Statistical Machine Translation Proceedings*. Lisboa, Portugal: Association for Computational Linguistics. 392–395.
12. Popović, M. 2016. chrF deconstructed: β parameters and n-gram weights. *1st Conference on Machine Translation Proceedings*. Berlin, Germany: Association for Computational Linguistics. 2:499–504.
13. Chi-kiu, Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. *Conference on Machine Translation Proceedings*. Copenhagen, Denmark: Association for Computational Linguistics. 2:589–597.
14. Stanojević, M., and K. Sima'an. 2014. BEER: BEtter evaluation as ranking. *9th Workshop on Statistical Machine Translation Proceedings*. Baltimore, MD: Association for Computational Linguistics. 414–419.
15. Stanojević, M., and K. Sima'an. 2015. Evaluating MT systems with BEER. *Prague Bulletin Mathematical Linguistics* 104:17–26.
16. Sellam, T., D. Das, and A. P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. Available at:

https://arxiv.org/pdf/2004.04696.pdf (accessed May 5, 2021).

17. Inkova, O. Yu. 2018. Nadkorpusnaya baza dannykh kak instrument formal'noy variativnosti konnektorov [Supra-corpora database as an instrument of the study of the formal variability of connectives]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po mat-lam ezhegodnoy Mezhdunar. konf. "Dialog"* [Computer Linguistic and Intellectual Technologies: Conference (International) "Dialog" Proceedings]. Moscow. 17(24):240–253.
18. Castilho, Sh., S. Doherty, F. Gaspari, and J. Moorkens. 2018. Approaches to human and machine translation quality assessment. *Translation quality assessment: From principles to practice*. Eds. J. Moorkens, Sh. Castilho, F. Gaspari, and S. Doherty. Cham: Springer. 9–38.
19. Likert, R. 1932. A technique for the measurement of attitudes. *Arch. Psychol.* 140:1–55.
20. Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76(5):378–382.
21. Shterionov, D., R. Superbo, P. Nagle, *et al.* 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* 32:217–235.
22. Castilho, S., J. Moorkens, F. Gaspari, *et al.* 2018. Evaluating MT for massive open online courses. A multifaceted comparison between PBSMT and NMT systems. *Machine Translation* 32:255–278.
23. Popovic, M. 2018. Error classification and analysis for machine translation quality assessment. *Translation quality assessment: From principles to practice*. Eds. J. Moorkens, Sh. Castilho, F. Gaspari, and S. Doherty. Cham: Springer. 129–158.
24. Lommel, A. 2018. Metrics for translation quality assessment: A case for standardising error typologies. *Translation quality assessment: From principles to practice*. Eds. J. Moorkens, Sh. Castilho, F. Gaspari, and S. Doherty. Cham: Springer. 109–127.
25. Klubička, F., A. Toral, and V. M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian. *Machine Translation* 32:195–215.
26. Haque, R., M. Hasanuzzaman, and A. Way. 2020. Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation* 34:149–195.
27. Vilar, D., J. Xu, L. D'Haro, and H. Ney. 2006. Error analysis of statistical machine translation output. *5th Conference (International) on Language Resources and Evaluation Proceedings*. 697–702.
28. Goncharov, A. A., N. V. Buntman, and V. A. Nuriev. 2019. Oshibki v mashinnom perevode: problemy klassifikatsii [Machine translation errors: Problems of classification]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(3):92–103.
29. Calixto, I., and Q. Liu. 2019. An error analysis for image-based multi-modal neural machine translation. *Machine Translation* 33:155–177.

Received April 14, 2021

Contributors

Nuriev Vitaly A. (b. 1980) — Candidate of Science (PhD) in philology, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; nurieff.v@gmail.com

Egorova Anna Yu. (b. 1991) — junior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; ann.shurova@gmail.com