УДК 81’322.4
DOI: 10.18384/2310-712X-2022-1-47-59

АВТОМАТИЧЕСКАЯ ОЦЕНКА КАЧЕСТВА МАШИННОГО ПЕРЕВОДА НАУЧНОГО ТЕКСТА: 5 ЛЕТ СПУСТЯ

Светлой памяти моего учителя
Нелюбина Л. Л.

Улиткин И. А.
Московский областной государственный университет
141014, Московская область, г. Мытищи, ул. Веры Волошиной, д. 24, Российская Федерация

Аннотация
Цель. Проведено сравнение переводов систем нейронного машинного перевода Google и PROMT с переводами, полученными 5 лет назад, когда в основе данных систем использовались алгоритмы статистического перевода и перевода на основе правил (см. Вестник МГОУ. Серия Лингвистика. 2016. № 4. С. 174–182).
Процедуры и методы. Использованы современные метрики автоматической оценки качества машинного перевода BLEU, F-measure и TER для сравнения качества переводов современных систем нейронного машинного перевода на примере онлайн-систем Google и PROMPT.
Результаты. Проведённая оценка качества перевода текстов-кандидатов Google и PROMT в сравнении с референтным переводом при помощи автоматической программы позволила выявить существенные качественные изменения по сравнению с результатами, полученными 5 лет назад, что свидетельствует о значительном улучшении работы вышеуказанных переводческих онлайн-сервисов.
Теоретическая и/или практическая значимость. Описанные способы автоматической оценки качества машинного перевода (МП), т. е. методы, основанные на сравнении строк, и *n*-граммные модели позволяют провести оценку качества переводов систем машинного перевода. Обсуждаются способы улучшения качества МП. Показано, что современные системы автоматической оценки качества перевода позволяют выявлять и систематизировать ошибки, допущенные системами МП, что позволит в будущем совершенствовать данные системы.

Ключевые слова: автоматическая оценка, качество перевода, машинный перевод, метрики, BLEU, TER, F-measure, эталонный перевод

AUTOMATIC EVALUATION OF MACHINE TRANSLATION QUALITY OF A SCIENTIFIC TEXT: FIVE YEARS LATER

I. Ulitkin
Moscow Region State University
24 ulitsa Very Voloshinoi, 141014 Mytishchi, Moscow region, Russian Federation

Abstract
Aim. The paper compares translations of Google and PROMT neural machine translation systems with translations obtained 5 years ago, when statistical machine translation and rule-based machine

translation algorithms were used, respectively, as the main translation algorithms of these systems (see Bulletin of Moscow Region State University. Series: Linguistics. 2016. no. 4. pp. 174-182).
Methodology. Use is made of such modern metrics for automatic quality evaluation of machine translation as BLEU, F-measure and TER to compare the quality of translations of modern neural machine translation systems using the example of Google and PROMPT online systems.
Results. The evaluation of the translation quality of candidate texts generated by Google and PROMT in comparison with the reference translation using an automatic translation evaluation program reveals significant qualitative changes as compared with the results obtained 5 years ago, which indicates a dramatic improvement in the work of the above-mentioned online translation systems.
Research implications. The described three methods for evaluating the quality of machine translation allow one to analyze several automatic methods for evaluating machine translation quality, i.e. methods based on string matching and *n*-gram models. Ways to improve the quality of machine translation are discussed. It is shown that modern systems of automatic translation quality evaluation allow errors made by machine translation systems to be identified and systematized, which will make it possible to improve the quality of translation by these systems in the future.

Keywords: automatic evaluation, translation quality, machine translation, metrics, BLEU, TER, F-measure, reference translation

1. Введение

Машинный перевод – это перевод текста с исходного языка в текст на языке перевода при помощи компьютерных программ; при этом профессиональные переводчики могут быть вовлечены на этапах предварительного редактирования исходного текста или последующего редактирования текста перевода, но обычно не участвуют в самом процессе перевода.

Хотя концепцию машинного перевода можно проследить до XVII века, но именно в 1950-х годах финансируемые правительством США исследования стимулировали интерес международного сообщества к исследованию и производству систем машинного перевода.

Первоначально предполагалось создать полностью автоматическую систему высококачественного машинного перевода (fully automated high-quality machine translation system), но к 1952 г. стало «ясно, что создание полностью автоматизированных систем мало реалистично, и что перевод с помощью таких систем будет требовать непосредственного участия человека» (перевод наш – *И. У.*) [10, p. 376].

Системы машинного перевода прошли долгий путь развития: от прямого перевода середины 50-х годов прошлого столетия к интерлингвальному подходу, который так и остался только на уровне идеи [13], затем к трансферному подходу, основанному на правилах, и к переводу с использованием предварительного редактирования текста [9], к статистическим системам машинного перевода и, наконец, к нейронному машинному переводу, анонсированному компанией Google 15 ноября 2016 г.

Различные исследователи систем машинного перевода неоднократно подчёркивали недостатки таких систем, выделяя в качестве основных медлительность, отсутствие точности и дороговизну [11], а также невозможность реализации знаний об окружающем нас мире через системы машинного перевода [5, p. 173]. Тем не менее ряд исследователей подчёркивает и перспективность подобных разработок. Так, в своей работе 1992 года А. Гросс продемонстрировал, что общие переводы, требующие знания реального мира, лучше переводятся человеком, в то время как тексты с математическими и абстрактными понятиями наиболее качественно переводятся системами машинного перевода [8, p. 103].

С ростом объёма информации изменилось отношение компаний к перево-

дам, поскольку их целью часто является простой обмен информацией. Например, работникам из Европейской комиссии часто требуется только представление о содержании документа. Им необходимо понять, стоит ли переводить текст в дальнейшем. А простые пользователи могут довольствоваться бесплатными интернет-сервисами машинного перевода, чтобы понять суть того, что написано на веб-сайте. То есть, когда существует потребность в простом понимании содержания текста, машинный перевод часто оказывается гораздо более быстрым и экономически эффективным решением, чем перевод, выполненный человеком.

Последние разработки в области машинного перевода привели к внедрению глубокого обучения и нейронных сетей для повышения точности переводов. Поставщики языковых услуг теперь предлагают индивидуальные механизмы машинного перевода, где помимо включения терминологии из определённой области, такой, как науки о жизни, туристической индустрии или информационных технологий, пользователь также может загрузить и использовать свою собственную память переводов (базы данных, содержащие ранее переведённые тексты), чтобы попытаться повысить точность, а также улучшить стиль и качество машинного перевода.

Всё вышесказанное свидетельствует о несомненном улучшении качества машинного перевода, что, прежде всего, связано с развитием технологий, доступностью больших параллельных корпусов для тренировки систем, а также огромным опытом, накопленном в области МП в последние десятилетия.

В настоящее время фиксируется стабильный рост использования различных онлайн-сервисов перевода текстов. На разработку подобных сервисов многие страны тратят огромные средства в надежде на то, что в скором будущем системы автоматического перевода смогут удовлетворить растущие потребности человека в переводе текстов различной жанровой направленности. Существующие на сегодняшний день автоматические метрики оценки качества перевода также постоянно совершенствуются и позволяют проводить объективную оценку качества переводов тех или иных систем.

Таким образом цель данной работы – анализ качества машинного перевода научно-технического текста, выполненного системами нейронного машинного перевода Google и PROMT, и сравнение полученных результатов с данными наших исследований в 2016 г. [3].

2. Методы оценки качества перевода

Качество перевода (под этим термином мы понимаем уровень качества выполненного письменного или устного перевода, оцениваемый в соответствии с рядом объективных и субъективных критериев, принятых для оценки данного вида письменного или устного перевода) зависит от целого ряда объективных и субъективных факторов. Оно определяется, прежде всего, **качеством исходного текста** или переводимой устной речи, а также **квалификацией переводчика** и его подготовленностью к осуществлению конкретного акта перевода.

Говоря о методах оценки качества перевода, выделяют экспертную оценку перевода (или ручную оценку качества перевода) и автоматическую.

2.1. Экспертная оценка перевода

«Как можно заметить, экспертная оценка профессионального перевода является довольно субъективной. Конкретный текст и задача перевода во многом определяют критерии и, как следствие, результат экспертной оценки» [1, с. 65].

Параметры, по которым эксперты оценивают машинный перевод и используемые при этом методы, как и в случае с экспертной оценкой профессионального перевода, варьируются в зависимости от проекта.

Например, по словам О. В. Митрениной, «в качестве ключевых параметров

могут выступать полнота (adequacy), которая оценивает точность перевода, и гладкость (fluency), отвечающая за правильность перевода» [2, с. 185].

О. В. Митренина также отмечает, что «другим способом оценки может быть выбор лучшего из предложенных вариантов перевода или ранжирование всех представленных вариантов. Эксперт также может оценивать перевод с точки зрения потраченных на перевод времени и сил, т. е. оценивать перевод по затраченным переводчиком-редактором ресурсам на исправление и доработку машинного перевода» [2, с. 185].

Первыми методиками экспертной оценки принято считать методики комитета ALPAC и APRA.

Консультативный комитет по автоматической обработке языков (ALPAC – Automatic Language Processing Advisory Committee) был создан в апреле 1964 г. «для оценки прогресса в компьютерной лингвистике вообще и в машинном переводе в частности. Пожалуй, самой известной работой комитета является отчёт, опубликованный в 1966 г. В нём подчеркиваются недостатки проведённых исследований в области машинного перевода, обращается внимание на необходимость фундаментальных исследований в области компьютерной лингвистики. Кроме того, в нём рекомендовалось прекратить государственное финансирование данной области исследований. В качестве основной причины указывались неудовлетворительные результаты, которые были получены за 10 лет исследований» [11].

Также в докладе ALPAC отмечалось, что «в основе экспертной оценки переводов лежит сравнение машинного перевода текста с русского на английский с эталонным человеческим переводом. При этом использовались следующие показатели: intelligibility (условная понятность, которую оценивали по шкале от 1 до 9) и fidelity / accuracy (точность перевода, которую можно было оценить от 0 до 9)» [11, p. 67].

Оценка 1 по шкале intelligibility даётся предложению, которое было непонятным, и даже контекст не помогает определить смысл. Оценка 9 ставится понятному предложению, которое не содержит стилистических ошибок.

Точность измеряется косвенно и отражает меру того, насколько информативно переведённое предложение по сравнению с оригиналом. Таким образом, по шкале от 0 до 9 баллов оценивается информативность (informativeness): 9 баллов по шкале точности характеризуют перевод как высокоинформативный, тогда как 0 баллов означает, что оригинал содержит меньше информации, чем перевод. Получается, что в случае с intelligibility оценивались переводные предложения, а по шкале fidelity оценивали оригинальные предложения. Можно сделать вывод, что если предложение оценивается в 9 баллов, то оно очень информативно.

Эксперты ALPAC пришли к следующим выводам: «Во-первых, усреднённые показатели понятности и точности являются сильно взаимосвязанными. Во-вторых, стало ясно, что минимальное количество экспертов должно составлять 4 человека. И, в-третьих, эксперты должны знать предметную область и язык оригинала для того, чтобы успешно оценивать перевод» [11, p. 73].

Управление перспективных исследовательских проектов Министерства обороны США (Defense Advanced Research Projects Agency (DARPA)), первоначально известное как Агентство перспективных исследовательских проектов (ARPA), было создано в феврале 1958 г. В 1991 г. в DARPA были проведены тесты статистических систем, основанных на правилах, и систем, требующих участия человека. В 1992 г. зарекомендовавшие себя методы были включены в стандартную программу тестирования.

Можно заключить, что экспертная оценка профессионального и машинного перевода требует большого труда и непосредственного участия людей. Эксперты,

которых привлекают к оценке перевода, могут проработать ограниченный объём предложений и текстов. Остаётся неясным, как оценивать перевод, какие определения давать и какие критерии применять.

2.2. Автоматическая оценка качества перевода

В основе автоматической оценки качества МП с использованием эталонных текстов лежит применение различных метрик, которые позволяют упростить и удешевить оценку качества.

Первая метрика, доступная для пары русский ↔ английский языки, – это BLEU (Bilingual Evaluation Understudy). Заметим, что данная метрика является одной из самых популярных и доступных на данный момент. Рассчитать значения BLEU можно также с помощью таких инструментов, как: MT-ComparEval, Interactive BLEU score evaluator и т. д.

Главная идея, положенная в основу данной метрики, заключается в следующем: чем ближе МП к переводу, выполненному профессиональным переводчиком, тем лучше. В ходе оценки качества машинного перевода, измеряется степень близости МП к одному или нескольким переводам, выполненным человеком, при помощи числовой метрики. Таким образом, указанная система оценки МП предполагает два компонента: 1) числовая метрика, по которой рассчитывается близость переводов, и 2) примеры (корпус) переводов хорошего качества, выполненных переводчиками [6, p. 311].

Метрики BLEU Score сравнивают *n*-граммы из перевода-кандидата с эталонным переводом, также производится подсчёт совпадений. Чем больше совпадений, тем лучше качество перевода-кандидата. При расчёте метрики BLEU важное значение приобретает количество переводов-эталонов: чем больше эталонов, тем точнее показатель качества перевода.

У метрики BLEU есть две составляющие. Первая - это точность или precision. Чтобы определить точность, подсчитывается количество тех слов (униграмм) из перевода-кандидата, которые встречаются в любом из переводов-эталонов. К сожалению, системы машинного перевода могут в некоторых случаях генерировать слишком большое количество «нужных» слов (результатом может, например, служить появление в переводе повторяющего артикля «the the the»), что, в свою очередь, может привести к слишком высоким показателям точности. Во избежание данной проблемы подсчитывается максимальное число слов из перевода-кандидата, которые есть в одном из эталонных переводов. Затем общее число слов каждого перевода-кандидата сводится к максимальному числу таких же (совпавших) слов в переводе-эталоне и делится на общее (неограниченное) число слов в переводе-кандидате [6, p. 312]. «Нужно отметить, что такой подсчёт происходит не только для униграмм, но для *n*-грамм. Такой расчёт точности даёт представление о двух аспектах перевода: адекватности (adequacy) и беглости (fluency). Перевод с использованием одинаковых слов (униграмм), что и в эталонном переводе, имеет тенденцию соответствовать адекватному переводу. Более длинные совпадения *n*-грамм говорят о беглости перевода» [6, p. 313].

Точность определяется путём перемножения всех *n*-грамм с последующим извлечением из произведения корня четвёртой степени – так получается среднее геометрическое.

Вторая составляющая метрики BLEU – это штраф за длину перевода или Brevity Penalty. Вычисление данного штрафа (BP) происходит следующим образом: BP равно 1, если длина перевода-кандидата больше длины перевода-эталона. BP меньше 1, если длина перевода-кандидата равна или меньше длины перевода-эталона [6, p. 315].

Особенностью метрики BLEU является то, что она основывается на точном совпадении форм слов. Можно утверждать, что применение данной метрики целесо-

образно для английского языка, где формы могут совпадать во многих случаях, однако не так удобно для русского языка. Важно и то, что в BLEU не учитывается синтаксис и порядок слов (но определяются более длинные совпадающие *n*-граммы).

Ещё одна мера, доступная для языковой пары русский ↔ английский языки, – это TER (Translation Edit Rate). TER рассчитывает количество исправлений, которые нужны для того, чтобы получившийся перевод семантически походил на эталонный. Эта мера сохраняет время и деньги. В стремлении добиться более высоких корреляций с существующими экспертными оценками, исследователи назначают более низкие оценки за фазовые сдвиги сочетаний (phrasal shifts) по сравнению с теми, которые назначают подходы, основанные на *n*-граммах (такие, как BLEU) [4, p. 231].

При использовании метрики TER определяется минимальное количество исправлений, которое необходимо внести в переведённый текст, чтобы он совпадал с референтным. Здесь измеряют только количество редактирований, поэтому для достижения эталонности перевода рассчитывается их минимальное количество.

Возможны следующие варианты редактирования текста: вставка, удаление и замена отдельных слов, а также сдвиги последовательностей слов. Сдвиг происходит в пределах перевода предложения, перемещением смежной последовательности слов. Все изменения, включая сдвиги любого количества слов на любое расстояние, имеют одинаковую «стоимость». Кроме того, знаки пунктуации рассматриваются как обычные слова, а неверное использование регистра считается редактированием [4, p. 223].

Существует также метрика F-measure; её разработчики утверждают, что именно она показывает наилучшее совпадение с оценкой, выполненной человеком [12]. Однако это не всегда так. Метрика F-measure не очень хорошо работает с небольшими отрезками [7].

3. Автоматическая оценка качества перевода систем нейронного машинного перевода Google и PROMT

Для проведения анализа в 2016 г. были отобраны 500 предложений из научных статей журнала «Квантовая электроника»[1] и их переводы на английский язык, выполненные профессиональными переводчиками. В том же году (2016 г.) русские предложения были переведены системами МП Google и PROMT, и эти переводы сравнивались с эталонным переводом[2]. Те же самые русские предложения были переведены на английский язык в 2021 г. при помощи систем МП Google и PROMT и снова проанализированы для оценки их качества.

В ходе автоматического анализа использовалась программа Language Studio™ Lite (размещённая на сайте http://www.languagestudio.com), которая позволяет оценить качество МП при помощи таких популярных метрик, как BLEU, F-Measure, TER [14].

3.1. Оценка качества перевода с помощью n-граммных метрик

Вначале мы сравнили референтный текст (под референтным текстом подразумевается выполненный переводчиком перевод) и тексты-кандидаты Google и PROMT (под текстами-кандидатами подразумеваются переводы, выполненные системами МП) при помощи *n*-граммной метрики. Результаты, полученные для онлайн переводчиков Google и PROMPT в 2016 г. и в 2021 г., представлены в табл. 1 и 2.

[1] «Квантовая электроника» – ведущий российский научный ежемесячный журнал в области лазеров и их применений, а также по связанным с ними тематикам. См.: Квантовая электроника [Электронный ресурс]. URL: http://www.quantum-electron.ru/ (дата обращения: 20.10.2021).

[2] Подробнее см.: Улиткин И. А. Автоматическая оценка качества машинного перевода научно-технического текста // Вестник Московского государственного областного университета. Серия: Лингвистика. 2016. № 4. С. 174–182.

Таблица 1 / Table 1

Анализ переводов онлайн-переводчика Google за 2016 и 2021 годы, на основе *n*-граммной метрики / Analysis of translations of Google Translate for 2016 and 2021, based on the *n*-gram model

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:20 AM	2/9/2021 11:40 AM
Job End Date:	12/29/2015 10:20 AM	2/9/2021 11:41 AM
Job Duration:	0 min(s) 12 sec(s)	0 min(s) 17 sec(s)
Reference File:	science_reference_corrected.txt	science_reference_corrected.txt
Candidate File:	science_google_corrected.txt	google_translation_2021.txt
Evaluation Lines:	500	500
Tokenization Language:	EN	EN
Results Summary:	**46.147**	**50.194**

Таблица 2 / Table 2

Анализ переводов онлайн переводчика Prompt за 2016 и 2021 годы на основе *n*-граммной метрики / Analysis of translations of PROMT translation service for 2016 and 2021, based on the *n*-gram model

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:21 AM	2/9/2021 11:43 AM
Job End Date:	12/29/2015 10:21 AM	2/9/2021 11:43 AM
Job Duration:	0 min(s) 12 sec(s)	0 min(s) 13 sec(s)
Reference File:	science_reference_corrected.txt	science_reference_corrected.txt
Candidate File:	science_PROMT_corrected.txt	prompt_translation_2021.txt
Evaluation Lines:	500	500
Tokenization Language:	EN	EN
Results Summary:	**30.791**	**44.420**

Полученные результаты показывают, что за последние 5 лет система МП Google улучшила свои показатели на 4% (50,19% совпадений в 2021 г. по сравнению с 46,14% в 2016 г.), что позволяет сделать вывод о несомненном росте качества машинного перевода на основе нейронного обучения. При переводе научных текстов система МП PROMT продемонстрировала почти 14%-ное улучшение качества перевода (44,42% совпадений в 2021 г. по сравнению с 30,79% в 2016 г.), что не удивительно, поскольку в 2016 г. в основе перевода системы МП PROMT была модель, основанная на правилах, а не модель статистического перевода на основе *n*-граммов.

Из полученных данных можно сделать вывод, что системы МП Google и PROMPT адекватно справляются с научно-техническими текстами, где преобладает терминология и простые предложения. Так, порог совпадений на уровне 75% для обеих систем МП составляет более 100 предложений из 500. Такой процент совпадений обеспечивает минимальные затраты на постредактирование машинного перевода.

Анализ полученных данных выявляет следующую закономерность: в предложениях с совпадением от 100% до 70% встречаются одна–две ошибки в переводе; в предложениях с соотношением от 69% до 50% можно наблюдать три ошиб-

ки и более. При этом обе системы МП прекрасно справляются с простыми распространёнными и нераспространёнными предложениями, а также со сложносочинёнными предложениями. Системы демонстрируют хорошее «знание» научной терминологии.

Основные ошибки, которые мы обнаружили при сравнении референтных и машинных переводов, связаны с неправильной передачей аббревиатур. Системы машинного перевода не всегда способны правильно перевести специализированные сокращения, с чем, в свою очередь, легко справляется профессиональный переводчик, который специализируется в той или иной области. Таким образом, для достижения лучшего качества МП необходимо проводить расшифровку всех аббревиатур на протяжении всего текста. Ещё один пример ошибок в машинных переводах – неправильный порядок слов. Следует, однако, отметить, что таких ошибок становится всё меньше в системах МП после их соответствующего обучения.

Так, сравнение референтных переводов с машинными переводами позволяет выделить следующие повторяющиеся ошибки в обеих системах МП: отсутствие артиклей, ошибки в значении и выборе слова, нарушение порядка слов в предложениях.

Полученные данные указывают на то, что системы МП Google и PROMPT постоянно совершенствуются. Последнее наводит на мысль, что потенциал нейронных систем машинного перевода со временем будет улучшаться.

3.2. Оценка качества перевода при помощи сопоставительных метрик BLEU, F-measure и TER

Второй анализ был проведён с использованием таких метрик, как BLEU, F-measure и Translation Error Rate (TER). Осуществлялось сравнение сразу двух текстов-кандидатов с референтным переводом. Результаты исследований для 2016 и 2021 гг. представлены в табл. 3 и 4.

Как и в предыдущем тесте, система МП Google демонстрирует небольшой рост качества перевода при сопоставлении результатов анализа переводов, выполненных статистической системой МП и нейронной системой МП. Одновременно с этим, система МП PROMPT показывает значительное улучшение своих показателей по сравнению с 2016 г., что объясняется переходом от перевода на основе правил к переводу на основе нейронного обучения.

Результаты сравнения данных за 2021 г. показывают, что происходит выравнивание показателей систем МП Google и PROMPT; это связано, прежде всего, с использованием нейронного обучения систем МП переводу в обеих системах.

Системы МП на основе нейронного обучения учатся на огромных корпусах существующих переводов на разные языковые пары. В отличие от статистического подхода к переводу, поисковые алгоритмы которого интуитивно предпочитают использовать последовательности слов, являющиеся наиболее вероятными переводами исходных (что позволяет с высокой точностью генерировать правильную последовательность слов на целевом языке), нейронные системы машинного перевода не просто ищут соответствия слову и фразам, а тщательно изучают взаимоотношения между двумя языками. Анализ каждого сегмента текста позволяет современным системам понять его контекст, определяя значение каждого слова, которое необходимо перевести. В результате такого анализа системы нейронного машинного перевода подбирают необходимые грамматические структуры, правильно воспроизводя семантику и структуру текста перевода.

В результате анализа мы обнаружили следующую тенденцию. Современные нейронные системы машинного перевода демонстрируют аналогичные результаты, что объясняется использованием похожих алгоритмов при генерировании переводов.

Таблица 3 / Table 3

Оценка качества переводов, выполненных системами МП Google и Prompt в 2016 г., при помощи сопоставительных метрик BLEU, F-measure и TER[1] / Evaluation of the quality of translations made by Google and Prompt in 2016 using BLEU, F-measure and TER metrics

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:17 AM
Job End Date:	12/29/2015 10:18 AM
Job Duration:	0 min(s) 44 sec(s)
Number of Reference Files:	1
Number of Candidate Files:	2
Evaluation Lines:	500
Tokenization Language:	EN
Evaluation Metrics:	BLEU, F-Measure, TER (Inverted Score)

Results Summary

Candidate File:	**1**	**2**
BLEU Case Sensitive	**24.54**	**42.10**
BLEU Case Insensitive	**25.98**	**43.62**
F-Measure Case Sensitive	**60.01**	**72.26**
F-Measure Case Insensitive	**61.35**	**73.24**
TER Case Sensitive	**38.07**	**54.43**
TER Case Insensitive	**38.70**	**54.94**

Candidate Files:

1 : science_PROMT_corrected.txt

2 : science_google_corrected.txt

Reference Files:

1 : science_reference_corrected.txt

-- Report End –

[1] См. также: Улиткин И. А. Автоматическая оценка качества машинного перевода научно-технического текста [3].

Таблица 4 / Table4

Оценка качества переводов, выполненных системами МП Google и Prompt в 2021 г., при помощи сопоставительных метрик BLEU, F-measure и TER / Evaluation of the quality of translations made by Google and Prompt in 2021 using BLEU, F-measure and TER metrics

Translation Evaluation Summary

Job Start Date:	2/9/2021 11:40 AM
Job End Date:	2/9/2021 11:41 AM
Job Duration:	0 min(s) 54 sec(s)
Number of Reference Files:	1
Number of Candidate Files:	2
Evaluation Lines:	500
Tokenization Language:	EN
Evaluation Metrics:	BLEU, F-Measure, TER (Inverted Score)

Results Summary

Candidate File:	**1**	**2**
BLEU Case Sensitive	**40.25**	**45.79**
BLEU Case Insensitive	**41.53**	**47.35**
F-Measure Case Sensitive	**71.41**	**75.05**
F-Measure Case Insensitive	**72.20**	**75.79**
TER Case Sensitive	**53.03**	**56.82**
TER Case Insensitive	**53.42**	**57.21**

Candidate Files:

1 : prompt_translation_2021.txt
2 : google_translation_2021.txt

Reference Files:

1 : science_reference_corrected.txt

-- Report End –

Метрика F-measure, основанная на поиске максимального количества соответствий между МП и референтными переводами (отношение между общим числом совпадающих слов к длине перевода и референтного текста), показывает наилучшие результаты. Это говорит о том, что в большинстве случаев количество слов в референтных текстах и текстах-кандидатах близко (более 70% для научных текстов при использовании систем МП Google и PROMT). Помимо этого совпадение идёт не только на уровне количества слов, но и на уровне лексики, что непосредственно связано с трудозатратами редактора, поскольку чем меньше ему придётся править текст, тем лучше.

Метрика TER, основанная на подсчёте количества поправок, показала результат хуже. Для научно-технических текстов –

более 50% при использовании МП Google и PROMT.

Наихудший результат из трёх метрик показала BLEU, основанная на *n*-граммах. Использование метрики BLEU позволяет определить, сколько слов совпадает в строке, и при этом наилучший результат дают не просто совпадающие слова, а последовательность слов. Для научно-технических текстов результат составил более 45% при использовании системы Google и более 40% при использовании системы PROMT.

4. Заключение

В данной статье представлен обзор наиболее часто используемых сегодня метрик оценки МП. Как правило, данные метрики показывают хорошую корреляцию переводов-кандидатов с референтными переводами. При этом для всех метрик выделен важный недостаток: они не могут предоставить оценку качества МП на уровне смысла. Тем не менее они являются наиболее популярными системами автоматической оценки качества МП.

Сравнение результатов, полученных в 2021 г., показывает заметное улучшение качества нейронного машинного перевода систем Google и PROMT по сравнению с 2016 г., что вполне обосновано, поскольку новая технология МП даёт заметные преимущества по сравнению со статистической моделью и тем более с системой машинного перевода, основанного на правилах.

Проведённое в данной работе сравнение референтного перевода с переводами Google и PROMT позволяет сделать вывод, что наибольшее количество ошибок происходит на уровне семантики, т. е. понимания машиной исходного текста. Всё это говорит о том, что в настоящее время ещё не существует необходимых баз данных семантических конструкций, которые позволили бы избежать повторения подобных ошибок. Также стоит отметить, что системы МП испытывают немалые затруднения при переводе сложных грамматических, синтаксических и лексических конструкций. Здесь важно понимать, что адекватная и полная автоматическая оценка качества переводов позволяет выявлять и систематизировать не только ошибки систем МП, но и недостатки существующих программ МП, что в будущем поможет решить выявленные проблемы.

Анализ качества МП текстов-кандидатов Google и PROMT в сравнении с референтным переводом, проведённый при помощи *n*-граммной модели и различных метрик, показывает, что перевод Google демонстрирует наилучшее соответствие с референтным переводом на уровне лексики. Это вполне ожидаемо, поскольку обучение системы проводится с использованием большого количества параллельных корпусов текстов; при этом происходит улучшение перевода и на синтаксическом уровне, что, вероятнее всего, связано с улучшением технологии перевода. Сравнение результатов 2016 и 2021 гг. для PROMT выявило наиболее заметный рост всех показателей, что связано с переходом на нейронное обучение данного онлайн-сервиса. Отставание от своего конкурента Google можно объяснить тем, что PROMT – это гибридная система, использующая преимущества нейронного обучения и перевода на основе правил. Однако все преимущества данной системы раскрываются наиболее полно лишь при активной тренировке PROMT на больших двуязычных корпусах (от 50000 сегментов), что не всегда легко реализовать на практике.

Разработка эффективных и надёжных метрик оценки МП в последние годы активно исследуется. Одна из важнейших задач – выйти за рамки *n*-граммной статистики, продолжая при этом использовать полностью автоматический режим. Потребность в полностью автоматических метриках нельзя недооценивать, поскольку именно они обеспечивают наибольшую скорость развития и прогресса систем МП.

Статья поступила в редакцию 20.10.2021

ЛИТЕРАТУРА

1. Комиссаров В. Н., Коралова А. Л. Практикум по переводу с английского языка на русский. М.: Высшая школа, 1990. 127 с.
2. Митренина О. В. Машинный перевод // Прикладная и компьютерная лингвистика / ред. Николаев И. С., Митренина О. В., Ландо Т. М. М.: УРСС, 2016. С. 156–189.
3. Улиткин И. А. Автоматическая оценка качества машинного перевода научно-технического текста // Вестник Московского государственного областного университета. Серия: Лингвистика. 2016. № 4. С. 174–182. DOI: 10.18384/2310-712X-2016-4-174-182.
4. A Study of Translation Edit Rate with Targeted Human Annotation / Snover M., Dorr B., Schwartz R., Micciulla L., Machoul J. // Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. 2006. P. 223–231.
5. Austermьhl F. Electronic Tools for Translators. Manchester: St. Jerome, 2001. 202 p.
6. Bleu: A Method for Automatic Evaluation of Machine Translation / Papineni K., Roukos S., Ward T., Zhu W.-J. // ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002. P. 311–318. DOI: 10.3115/1073083.1073135.
7. Confidence Estimation for Machine Translation / Blatz J., Fitzgerald E., Foster G., Gandrabur S., Goutte C., Kulesza A., Sanchis A., Ueffing N. // COLING' 04. Proceedings of the 20th International conference on Computational Linguistics. 2004. P. 315–321. DOI: 10.3115/1220355.1220401.
8. Gross A. Limitations of Computers as Translation Tools // Computers and Translation / ed. J. Newton. London: Routledge, 1992. P. 96–130.
9. Hutchins J. Current commercial machine translation systems and computer-based translation tools: System types and their uses // International Journal of Translation. 2005. Vol. 17 (1-2). P. 5–38.
10. Hutchins J. Machine Translation: History // Encyclopedia of Language and Linguistics / ed. K. Brown. Oxford: Elsevier, 2006. P. 375–383.
11. Language and Machines: Computers in Translation and Linguistics: a report by the Automatic Language Processing Advisory Committee, National Academy of Science / Pierce J. R., Carroll J. B., Hamp E. P., Hays D. G., et al. Washington, DC: The National Academic Press, 1966. 138 p.
12. Melamed I. D., Green R., Turian J. P. Precision and Recall of Machine Translation // NAACL-Short '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion volume of the Proceedings of HLT-NAACL 2003 (short papers). 2003. Vol. 2. P. 61–63. DOI: 10.3115/1073483.1073504.
13. Quah C. K. Translation and Technology. Basingstoke: Palgrave, 2006. 221 p.
14. Ulitkin I. A. Human Translation vs. Machine Translation: Rise of the Machines // Translation Journal. 2013. Vol. 17. № 1 [Электронный ресурс]. URL: http://translationjournal.net/journal/63mtquality.htm (дата обращения: 20.10.2021).

REFERENCES

1. Komissarov V. N., Koralova A. L. *Praktikum po perevodu s angliiskogo yazyka na russkii* [Workshop on translation from English to Russian]. Moscow, Vysshaya shkola Publ., 1990. 127 p.
2. Mitrenina O. V. [Machine translation]. In: *Prikladnaya i komp'yuternaya lingvistika* [Applied and Computational Linguistics]. Moscow, URSS Publ., 2016, pp. 156–189.
3. Ulitkin I. A. [Automatic evaluation of machine translation quality of a scientific text]. In: *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika* [Bulletin of the Moscow Region State University. Series: Linguistics], 2016, no. 4, pp. 174–182. DOI: 10.18384/2310-712X-2016-4-174-182.
4. Snover M., Dorr B., Schwartz R., Micciulla L., Machoul J. 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: Technical Papers*, 2006, pp. 223–231.
5. Austermühl F. Electronic Tools for Translators. Manchester, St. Jerome, 2001. 202 p.
6. Papineni K., Roukos S., Ward T., Zhu W.-J. Bleu: A Method for Automatic Evaluation of Machine Translation. In: *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.
7. Blatz J., Fitzgerald E., Foster G., Gandrabur S., Goutte C., Kulesza A., Sanchis A., Ueffing N. Confidence Estimation for Machine Translation. In: *COLING' 04. Proceedings of the 20th International conference on Computational Linguistics*, 2004, pp. 315–321. DOI: 10.3115/1220355.1220401.

8. Gross A. Limitations of Computers as Translation Tools. In: Newton J., ed. *Computers and Translation*. London, Routledge, 1992. pp. 96–130.
9. Hutchins J. Current commercial machine translation systems and computer-based translation tools: System types and their uses. In: *International Journal of Translation*, 2005, vol. 17 (1-2), pp. 5–38.
10. Hutchins J. Machine Translation: History. In: Brown K., ed. *Encyclopedia of Language and Linguistics*. Oxford, Elsevier, 2006, pp. 375–383.
11. Pierce J. R., Carroll J. B., Hamp E. P., Hays D. G., et al. Language and Machines: Computers in Translation and Linguistics: a report by the Automatic Language Processing Advisory Committee, National Academy of Science. Washington, DC, The National Academic Press, 1966. 138 p.
12. Melamed I. D., Green R., Turian J. P. Precision and Recall of Machine Translation. In: *NAACL-Short '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion volume of the Proceedings of HLT-NAACL 2003 (short papers)*, 2003, vol. 2, pp. 61–63. DOI: 10.3115/1073483.1073504.
13. Quah C. K. Translation and Technology. Basingstoke, Palgrave, 2006. 221 p.
14. Ulitkin I. A. Human Translation vs. Machine Translation: Rise of the Machines. In: *Translation Journal*, 2013, vol. 17, no. 1. Available at: http://translationjournal.net/journal/63mtquality.htm (accessed: 20.10.2021).

ИНФОРМАЦИЯ ОБ АВТОРЕ

Улиткин Илья Алексеевич – кандидат филологических наук, доцент кафедры переводоведения и когнитивной лингвистики Института лингвистики и межкультурной коммуникации Московского государственного областного университета;
e-mail: ulitkin-ilya@yandex.ru
ORCID: 0000-0002-6523-1526. eLIBRARY SPIN-код: 2795-8060.

INFORMATION ABOUT THE AUTHOR

Ilya A. Ulitkin – Can. Sci. (Philology), Assoc. Prof., Department of Translation Studies and Cognitive Linguistics, Institute of Linguistics and Intercultural Communication, Moscow Region State University;
e-mail: ulitkin-ilya@yandex.ru
ORCID iD: 0000-0002-6523-1526. eLIBRARY SPIN-код: 2795-8060.

ПРАВИЛЬНАЯ ССЫЛКА НА СТАТЬЮ

Улиткин И. А. Автоматическая оценка качества машинного перевода научного текста: 5 лет спустя // Вестник Московского государственного областного университета. Серия: Лингвистика. 2022. № 1. С. 47–59.
DOI: 10.18384/2310-712X-2022-1-47-59

FOR CITATION

Ulitkin I. A. Automatic evaluation of machine translation quality of a scientific text: five years later. In: *Bulletin of the Moscow Region State University. Series: Linguistics*, 2022, no. 1, pp. 47–59.
DOI: 10.18384/2310-712X-2022-1-47-59