УДК 81’322.4
DOI: 10.18384/2310-712X-2016-4-174-182

АВТОМАТИЧЕСКАЯ ОЦЕНКА КАЧЕСТВА МАШИННОГО ПЕРЕВОДА НАУЧНО-ТЕХНИЧЕСКОГО ТЕКСТА

Улиткин И.А.
Московский государственный областной университет
105005, г. Москва, ул. Радио, д. 10А, Российская Федерация

Аннотация. Рассмотрены различные подходы к автоматической оценке качества машинного перевода (МП). Описаны три способа оценки качества машинного перевода. Проведен анализ нескольких автоматических методов оценки МП, т.е. методы, основанные на сравнение строк, и n-граммные модели. Проведена оценка качества перевода текстов-кандидатов Google и PROMT с референтным переводом при помощи автоматической программы. Обсуждаются способы улучшения качества машинного перевода.

Ключевые слова: автоматическая оценка, качество перевода, машинный перевод, метрики, эталонный перевод.

AUTOMATIC EVALUATION OF MACHINE TRANSLATION QUALITY OF A SCIENTIFIC TEXT

I. Ulitkin
Moscow State Regional University
10A, Radio str., Moscow, 105005, Russian Federation

Abstract. Different approaches to automatic evaluation of machine translation quality are considered. Three methods of machine translation quality evaluation are described. We analyze several methods for automatic evaluation of MT, i.e., based on string matching and n-gram models. The candidate translations made by Google and PROMT are compared with the reference translation by an automatic translation evaluation program and the results of evaluation are presented. Ways of improving machine translation quality are discussed.

Key words: automatic evaluation, quality of translation, machine translation, metrics, reference translation.

Количество людей, использующих сегодня в своей деятельности возможности онлайн-сервисов перевода текстов с одного языка на другой, неуклонно растет. Помимо этого, для автоматизации процесса перевода широко используются и компьютерные программы, на разработку которых страны тратят миллионы долларов в надежде, что в будущем эти программы автоматического перевода будут способны удовлетворить растущие потребности человека в переводе

с одного языка на другой. Вместе с развитием систем машинного перевода появилась необходимость разработки и развития надежных способов оценки качества машинного перевода.

Качество перевода (под этим термином мы понимаем качественный уровень выполненного письменного или устного перевода, оцениваемый исходя из некого набора 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 профессиональной **квалификацией переводчика** и его подготовленностью к осуществлению данного конкретного акта перевода.

Целью данной работы является рассмотрение различных методов оценки качества МП и анализ качества перевода научно-технического с опорой на эталонный (референтный) текст.

Методы оценки качества МП

2.1. Ручная оценка качества перевода

В основе данного метода лежит оценка качества машинного перевода человеком по определенным критериям и определенной шкале, например, грамотность и адекватность перевода с помощью человеческих сил.

Оценка качества машинного перевода ручным способом является очень дорогим, субъективным и трудоемким процессом (Koehn and Monz, 2006; Turian et al., 2003). Кроме того, существует возможность, что асессор во время процесса будет уставать, проголодается, или ему может наскучить эта работа, что может отрицательно повлиять на его суждения [16, p. 219]. Тем не менее, автоматические метрики не могут заменить оценку, сделанную человеком. Уайт [16, p. 214] отмечает, что носители языка сразу могут интуитивно идентифицировать хорошее или плохое предложение, написанное на их родном языке. Им не надо думать об этом, анализировать предложение, или иметь большие лингвистические знания. Для оценки качества переводного текста, однако, важно рассматривать и исходный текст, и переводный текст. При этом предлагаются два параметра, которые и по сей день обычно применяются в условиях ручной оценки точности машинного перевода, то есть эквивалентность и точность.

Однако оценка качества перевода с точки зрения его эквивалентность и точности не является единственным способом. Каллисон-Берч с соавторами (Callison-Burch et al., 2007) считают, что при оценке текста можно использовать рейтинговую систему, т.е. асессоры должны оценить качество перевода по ряду критериев, т.е. составить рейтинг результатов разных систем машинного перевода и располагать их по порядку ((1): хуже всего до (5): лучше всего).

Третьим способом оценки качества перевода является анализ ошибок (Vilar et al., 2006; Stymne et al., 2012). Под анализом ошибок подразумевается идентификация и классификация ошибок в переводном тексте, полученном при машинном переводе.

Таким образом, ручная оценка качества слишком дорога, требовательна к асессорам и, зачастую, довольно несогласованна (оценка больше зависит от мнения асессора, нежели объективного качества).

2.2. Автоматическая оценка качества перевода без использования референтных текстов

Автоматическая оценка качества МП без использования эталонных текстов – новое направление в области оценки качества перевода, при котором осуществляется попытка отказаться от использования эталонов. Это изменение делает оценку качества практически бесплатной. Получение подобного инструмента могло было бы открыть следующие возможности (Bojara et al., 2013): 1) принятие решения, достаточно ли хорош перевод для публикации без постредактирования; 2) информирование читателя перевода о качестве перевода; 3) фильтрация предложений, которые не имеет смысла редактировать, а лучше перевести при помощи профессионального переводчика; 4) выбор лучшего перевода из нескольких вариантов.

Актуальность данной задачи подтверждается тем фактом, что созданием подобной метрики занимаются в рамках семинара по статистическому машинному переводу с 2012 г. (Callison-Burch et al., 2012). Каждый год формулировка конкретной задачи меняется, переводы пытаются оценивать как просто с помощью абстрактной оценки (1 – перевод идеальный, 2 – одна-две ошибки, 3 – ужасный), так и с помощью прикладных метрик (HTER – сколько шагов потребуется сделать переводчику, чтобы обработать полученный перевод для публикации (Bojara et al., 2013)).

2.3. Автоматическая оценка качества перевода с использованием референтного перевода

В основе автоматической оценки качества МП с использованием эталонных текстов лежит использование различных метрик, которые используются для упрощения и удешевления оценки качества.

Существует несколько автоматических методов оценки машинного перевода: одни из них основаны на сравнении строк (string matching), другие, например n-граммные модели, на использовании информационного поиска (information retrieval) [1, с. 61–77].

Яркими представителями методов на основе сравнения строк являются метрика пословной вероятности ошибок WER, метрика позиционно-независимой пословной вероятности ошибки PER и метрика вероятности ошибки перевода TER.

N-граммные методы оценки качества перевода представлены такими метриками как BLEU, NIST, METEOR и F-measure, которые основываются на подсчете точности n-грамм между эталоном и переводом некого текста.

Метрика BLEU получила свое распространение в таких видоизменнных метриках как Smoothed Bleu или BleuS (Lin et al., 2004) и NIST (National Institute of Standards and Technology) [6, pp. 128–132]. Первые используются для вычисления совпадений на уровне предложений и при этом довольно неплохо коррелируются с ручными оценками в статистическом машинном перевода, но при этом имеют проблемы с другими видами перевода [7]. В NIST используется также частотная составляющая (полнота и точность). Если BLEU просто вычисляет точность n-граммов, добавляя равный вес за каждое точное совпадение, NIST вычисляет также, насколько информативен каждый совпадающий n-грамм.

Разработчики F-measure утверждают, что именно их метрика показывает наилучшее совпадение с оценкой, выполненной человеком [11, pp. 61–63]. Однако это не всегда так. Метрика F-measure не очень хорошо работает с небольшими отрезками [2, pp. 315–321].

В России выпускаются программы для оценки автоматизированного перевода. Компания ПРОМТ 1 июля 2003 года выпустила программу Corvet, в которую заложены функции сравнения результатов машинного перевода текста и перевода после «ручной» обработки (некоторого «идеального перевода», например, сохраненного в формате TRADOS Translation Memory). Программа также позволяет сравнить качество вариантов перевода, выполненных разными людьми-переводчиками, или разными системами перевода. Программа Corvet программа сможет оказать неоценимую помощь пользователям, находящимся в процессе поиска и выбора корпоративного решения для автоматизации процесса перевода (только системы МП, только системы ТМ, или комплексное решение на основе интеграции этих систем). Выбранное на основе полученных от программы результатов решение будет гарантией эффективного достижения поставленных целей в каждом конкретном случае [http://www.PROMT.ru].

Автоматическая оценка качества перевода статистических и традиционных систем Google и PROMT

Перевод текста – задача интеллектуальная, поэтому скепсис в отношении возможности использования для этой цели компьютера вполне закономерен. Однако создателям систем МП удалось, что называется, наделить разумом свои разработки, и сейчас машинный перевод относят к классу технологий искусственного интеллекта.

Сегодня в мире существует два подхода к построению технологии машинного перевода: традиционный, который предполагает построение системы перевода на основе лингвистических правил (rule-based machine translation), и статистический (statistical-based machine translation), полагающийся на математические модели при обработке текста.

Для проведения анализа, мы отобрали 500 предложений из научных статей журнала «Квантовая электроника» (http://www.quantum-electron.ru/) и их переводы на английский язык, выполненные профессиональными переводчиками.

Для проведения автоматического анализа мы воспользовались программой Language StudioTM Lite с сайта (http://www.languagestudio.com), которая является бесплатной и позволяет оценить качество МП при помощи таких популярных метрик, как BLEU, F-Measure, TER.

3.1. Оценка качества перевода при помощи 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 выполненные МП системами переводы) при помощи n-граммной метрики.

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:20 AM
Job End Date:	12/29/2015 10:20 AM
Job Duration:	0 min(s) 12 sec(s)
Reference File:	science_reference_corrected.txt
Candidate File:	science_google_corrected.txt
Evaluation Lines:	500
Tokenization Language:	EN

Results Summary: 46.147

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:21 AM
Job End Date:	12/29/2015 10:21 AM
Job Duration:	0 min(s) 12 sec(s)
Reference File:	science_reference_corrected.txt
Candidate File:	science_PROMT_corrected.txt
Evaluation Lines:	500
Tokenization Language:	EN

Results Summary: 30.791

Полученные результаты показывают, что лидирует система МП Google, которя показывает наилучшее совпадение (46.14 %) при переводе научных текстов по сравнению с системой МП PROMT (30.791 %), что не удивительно, поскольку основу статистического перевода составляет n-граммная модель, преимущества которой проявляются при достаточно долгой тренировке на большом количестве корпусов текста.

При условии, что перевод на английский для Google является приоритетным, данная система МП постоянно совершенствуется. Последнее наводит на мысль, что потенциал трансфертных систем перевода рано или поздно будет исчерпан, в то время как качество перевода статистических систем МП со временем будет улучшаться.

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

Второй анализ был проведен с использованием таких метрик как BLEU, F-measure и Translation Error Rate (TER). Осуществлялось сравнение сразу двух текстов-кандидатов с референтным переводом. В результате мы получили следующие показатели:

Переводы научно-технических текстов

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 –

Как и в предыдущем тесте, система МП Google показывает более высокие результаты, что неудивительно, поскольку основой языкового оформления научных текстов является стандартизованность, то есть выбор предписываемого для данных условий коммуникации клишированного языкового варианта.

В результате анализа мы обнаружили следующую тенденцию. Будучи основанной на поиске максимального количества соответствий между системами МП и рефентными переводами, то есть отношение между общим числом совпадающих слов к длине перевода и референтного текста, метрика F-measure показывает наибольшие

результаты. Это говорит о том, что по большей части количество слов в референтных текстах и текстах-кандидатах близко (более 70 % для научных текстов при использовании системы МП Google и более 60 % при использовании системы PROMT). Помимо этого совпадение идет не только на уровне количества слов, но и на уровне лексики, что также достаточно важно, поскольку чем меньше придется редактору править текст тем лучше.

Метрика TER основанная на измерении количества поправок показала результат хуже. Для научно-технических текстов более 50% при использовании МП Google и более 30 % при использовании системы PROMT.

Наихудший результат из трех метрик показала BLEU, основанная на n-граммах. Метрика BLEU определяет сколько слов совпадает в строке, и при этом наибольший результат дают не просто совпадающие слова, а последовательность слов. Для научно-технических текстов результат составил более 40% при использовании системы Google и более 20% при использовании системы PROMT.

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. Одним из важных недостатков всех этих метрик является то, что они не могут предоставить оценку качества МП на уровне смысла. Тем не менее на данный момент они являются единственными системами автоматической оценки качества МП.

Проведен анализ качества МП текстов-кандидатов Google и PROMT с референтным переводом при помощи n-граммной модели и различных метрик. В обоих случаях, перевод Google показывает хорошее соответствие с референтным переводом. Наилучшее совпадение зарегистрировано на уровне лексики, что вполне ожидаемо, поскольку основу статистического перевода Google составляет n-граммная модель. Наихудший результат с точки зрения грамматики также демонстрирует Google, что также понятно, поскольку PROMT реализует RBMT-модель, в которой перевод зависит от объемов лингвистических баз данных (словарей) и глубины описания естественных языков, т.е. необходим учет максимального количества особенностей грамматической структуры.

Разработка эффективных и надежных 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.

ЛИТЕРАТУРА

1. Улиткин И.А. Автоматическая оценка качества машинного перевода // Перевод и когнитология в XXI веке: материалы V международной научной теоретической конференции (27-30 апреля 2012). М.: Изд-во МГОУ, 2012. С. 61–77.
2. Confidence estimation for machine translation // Proceedings of COLING (Geneva, Swit-

zerland, 2004) / Blatz J., Fitzgerald E., Foster G., Gandrabur S., Goutte C., Kulesza A., Sanchis A., Ueffing N. Geneva, 2004. C. 315–321.
3. Proc. VIII Workshop on Statistical Machine Translation / Bojar O., Buck C., Callison-Burch C., Federmann C., Haddow B., Koehn P., Monz C., Post M., Soricut R., Specia L. Sofia, 2013. C. 1–44.
4. Proc. Second Workshop on Statistical Machine Translation / Callison-Burch C., Fordyce C., Koehn P., Monz C., Schroeder J. Prague, 2007. C. 136–158.
5. Proc. VII Workshop on Statistical Machine Translation / Callison-Burch C., Koehn P., Monz C., Post M., Soricut R., Specia L. Montreal, 2012. C. 10.
6. Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics // Proceedings of the Human Language Technology Conference (HLT). San Diego, CA, 2002. C. 128–132.
7. Koehn P. Statistical Machine Translation . Cambridge: Cambridge University Press, 2010.
8. Koehn P., Monz C. Workshop on Statistical Machine Translation // Proc. NAACL 2006 . New York, USA, 2006. C. 102–121.
9. Lin C.-Y., Och F.J. Proc. Coling 2004 . Geneva, 2004.
10. Melamed I. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons // Third Workshop on Very Large Corpora (WVLC3). Boston, MA, 1995. C. 184–198.
11. Melamed I.D., Green R., Turian J.P. [Электронный ресурс]. URL: https://aclweb.org/anthology/N/N03/N03-2021.pdf] (дата обращения: 08.02.2016)
12. Stymne S., Ahrenberg L. [Электронный ресурс]. URL: http://www.mt-archive.info/LREC-2012- Stymne-2.pdf] (дата обращения: 10.08.2015)
13. Turian J.P., Shen L., Melamed I.D. Proc. MT Summit IX . New Orleans, USA, 2003.
14. Vilar D., Xu J., d'Haro L.F., Ney H. [Электронный ресурс]. URL: http://hnk.ffzg.hr/bibl/lrec2006/pdf/413_pdf.pdf] (дата обращения: 10.08.2015)
15. White J., O'Connell T., Carlson L. Evaluation of machine translation // In Human Language Technology: Proceedings of the Workshop (ARPA). Morristown, NJ, USA, 1993. C. 206–210.
16. White J.S. How to evaluate machine translation // Computers and Translation: A Translator's Guide / Somers, H. Amsterdam/Philadelphia: John Benjamins, 2003. C. 211–244.

REFERENCES

1. Ulitkin I.A. Avtomaticheskaya otsenka kachestva mashinnogo perevoda [Automatic evaluation of machine translation quality] Perevod i kognitologiya v XXI veke: materialy V mezhdunarodnoi nauchnoi teoreticheskoi konferentsii (27-30 aprelya 2012) [Translation and kognitologiya in the XXI century: materials of V international scientific theoretical conference (27-30 April 2012)]. M., Izd-vo MGOU, 2012. pp. 61–77.
2. Blatz J., Fitzgerald E., Foster G., Gandrabur S., Goutte C., Kulesza A., Sanchis A., Ueffing N. Confidence estimation for machine translation. Geneva, 2004. pp. 315–321.
3. Bojar O., Buck C., Callison-Burch C., Federmann C., Haddow B., Koehn P., Monz C., Post M., Soricut R., Specia L. Proc. VIII Workshop on Statistical Machine Translation. Sofia, 2013. pp. 1–44.
4. Callison-Burch C., Fordyce C., Koehn P., Monz C., Schroeder J. Proc. Second Workshop on Statistical Machine Translation. Prague, 2007. pp. 136–158.
5. Callison-Burch C., Koehn P., Monz C., Post M., Soricut R., Specia L. Proc. VII Workshop on Statistical Machine Translation. Montreal, 2012. pp. 10.
6. Doddington G. Automatic evaluation of machine translation quality using n-gram co-

occurrence statistics // Proceedings of the Human Language Technology Conference (HLT) [Proceedings of the Human Language Technology Conference (HLT)]. San Diego, CA, 2002. pp. 128–132.
7. Koehn P. Statistical Machine Translation. Cambridge, Cambridge University Press, 2010.
8. Koehn P., Monz C. Workshop on Statistical Machine Translation // Proc. NAACL 2006. New York, USA, 2006. pp. 102–121.
9. Lin C.-Y., Och F.J. Proc. Coling 2004. Geneva, 2004.
10. Melamed I. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons // Third Workshop on Very Large Corpora (WVLC3). Boston, MA, 1995. pp. 184–198.
11. Melamed I.D., Green R., Turian J.P. [Electronic resource]. URL: https://aclweb.org/anthology/N/N03/N03-2021.pdf] (request date: 08.02.2016)
12. Stymne S., Ahrenberg L. [Electronic resource]. URL: http://www.mt-archive.info/LREC-2012- Stymne-2.pdf] (request date: 10.08.2015)
13. Turian J.P., Shen L., Melamed I.D. Proc. MT Summit IX. New Orleans, USA, 2003.
14. Vilar D., Xu J., d'Haro L.F., Ney H. [Electronic resource]. URL: http://hnk.ffzg.hr/bibl/lrec2006/pdf/413_pdf.pdf] (request date: 10.08.2015)
15. White J., O'Connell T., Carlson L. Evaluation of machine translation // Human Language Technology: Proceedings of the Workshop (ARPA). Morristown, NJ, USA, 1993. pp. 206–210.
16. White J.S. How to evaluate machine translation // Computers and Translation: A Translator's Guide / Somers, H. Amsterdam/Philadelphia: John Benjamins, 2003. pp. 211–244.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Улиткин Илья Алексеевич – кандидат филологических наук, доцент, доцент кафедры переводоведения и когнитивной лингвистики Московского государственного областного университета;
e-mail: ulitkin-ilya@yandex.ru

INFORMATION ABOUT THE AUTHOR

Ilya Ulitkin – candidate of philological sciences, associate professor, associate professor of the department of translatology and cognitive linguistics of Moscow State Regional University;
e-mail: ulitkin-ilya@yandex.ru

БИБЛИОГРАФИЧЕСКАЯ ССЫЛКА

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

BIBLIOGRAPHIC REFERENCE

I. Ulitkin. Automatic evaluation of machine translation quality of a scientific text // Bulletin of Moscow State Region University. Series: Linguistics. 2016. no. 4. pp. 174-182.
DOI: 10.18384/2310-712X-2016-4-174-182