Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и управление

КАФЕДРА Автоматизированные системы обработки информации и управления

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Анализ англо-русского перевода. Сравнительный анализ двух характеристик: word error rate и bleu score.

Студент ИУ5-64Б — **Р.А. Низовцев**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель — **Ю.Е. Гапанюк**
(Подпись, дата) (И.О.Фамилия)

Консультант
(Подпись, дата) (И.О.Фамилия)

2022 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой __ИУ5____
(Индекс)

_____ _____
(И.О.Фамилия)

« _____ » _____ 20 ____ г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Анализ англо-русского перевода. Сравнительный анализ двух характеристик: word__ error rate и bleu score._____

_____

Студент группы ___ИУ5-64Б_____

Низовцев Роман Александрович_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

Учебная_____

Источник тематики (кафедра, предприятие, НИР) __НИР_____

График выполнения НИР: 25% к ___ нед., 50% к ___ нед., 75% к __ нед., 100% к ___ нед.

Техническое задание: произвести анализ англо-русского перевода с помощью двух характеристик: word error rate и bleu score. Сравнить эти характеристики.

_____

_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на __6__ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

_____

_____

_____

_____

Дата выдачи задания « 14 » _марта_____ 2022г.

Руководитель НИР _____ _Ю.Е. Гапанюк_____
(Подпись, дата) (И.О.Фамилия)

Студент _____ _Р.А. Низовцев_____
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

УДК 004.89

АНАЛИЗ АНГЛО-РУССКОГО ПРЕВОДА. СРАВНИТЕЛЬНЫЙ АНАЛИЗ ДВУХ ХАРАКТЕРИСТИК: WORD ERROR RATE И BLEU SCORE.

ANALYSIS OF THE ENGLISH-RUSSIAN TRANSLATION. COMPARATIVE ANALYSIS OF TWO CHARACTERISTICS: WORD ERROR RATE AND BLEU SCORE.

Низовцев Р.А.
Москва, МГТУ им. Н.Э. Баумана
Nizovtsev R.A.
Moscow, Bauman Moscow State Technical University

Аннотация. *В статье рассмотрено два способа определения качества англо-русского машинного перевода текстов. Конкретно рассматривается определение двух характеристик: word error rate и bleu scroe . Исходные данные взяты из примеров переводов агенства переводов «Лингваконтакт». Автор провел сравнительный анализ двух алгоритмов: определения word error rate и определения bleu score. Кроме того был проведен сравнительный анализ качества перевода трех лучших переводчиков, использующих технологии машинного обучения: Google, Yandex, Systran. Сравнение проводилось на трех типах текстов: простом, техническом и экономическом. Далее на основе уже обученных моделей была написана и протестирована программа для определения собственной характеристики, отражающей качество перевода в процентах. Для решения всех вышеперечисленных задач были задействованы библиотеки nltk, numpy и jiwer.*
Ключевые слова: *Word error rate, bleu score, машинный перевод, анализ перевода.*
Abstract. *The article considers two ways to determine the quality of English-Russian machine translation of texts. The definition of two characteristics is specifically considered: word error rate and bleu scroe . The source data is taken from the examples of translations of the translation agency "Linguacontact". The author conducted a comparative analysis of two algorithms: word error rate definitions and bleu score definitions. In addition, a comparative analysis of the translation quality of the three best translators using machine learning technologies was carried out: Google, Yandex, Systran. The comparison was carried out on three types of texts: simple, technical and economic. Then, based on the already trained models, a program was written and tested to determine its own characteristics reflecting the quality of the translation as a percentage. nltk, numpy and jiwer libraries were used to solve all the above tasks.*
Keywords: *Word error rate, bleu score, machine translation, translation analysis.*

Алгоритмы машинного перевода сегодня поражают своей сложностью и высокой точностью решения поставленной задачи. Они позволяют моментально переводить тексты, подбирая перевод слов в зависимости от контекста. Для оценки качества перевода проводится его сравнение с эталоном, после вычисляется коэфициент: word error rate или bleu score. Разберем каждую характеристику подробнее.

Word Error Rate

В последнее время в качестве основного показателя точности работы систем распознавания речи используется показатель WER, а именно, его абсолютное значение или относительное, если сравниваются различные модели/системы.

В соответствии с источником [2] метод определения показателя WER состоит в выравнивании двух текстовых строк (первая — это результат машинного перевода, а вторая – эталон перевода. Для подсчета WER используется расстояние Левенштейна. Оно представляет собой “стоимость“ редактирования данных (минимальное количество или взвешенная сумма операций редактирования) для преобразования первой строки во вторую с наименьшим числом операций ручной замены (*S*), удаления (*D*) и вставки (*I*) слов:
WER= (S+D+I)/T.

Иными словами, WER подсчитывает количество слов, неверно определенных во время распознавания, делит сумму на общее число слов, предоставленных в транскрибировании от человека (переменная N в следующей формуле), а затем умножает результат деления на 100, чтобы вычислить частоту ошибок в процентах.
Неправильно распознанные слова делятся на три категории:

- Вставка (I): слова, неправильно добавленные в расшифровку гипотезы
- Удаление (D): слова, не обнаруженные в расшифровке гипотезы
- Замена (S): слова, отличающиеся между эталоном и гипотезой

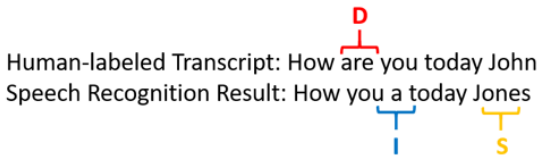

Рис.1 Принцип word error rate

Bleu Score:

В соответствии с источником [1] BLEU (Bilingual Evalution Understudy)— это измерение различий между автоматическим переводом и одним или несколькими эталонными пользовательскими переводами одного исходного предложения.

Алгоритм BLEU сравнивает последовательные фразы автоматического перевода с последовательными фразами, которые он находит в эталонном переводе, и взвешенно подсчитывает количество совпадений. Эти совпадения не зависят от позиции. Высшая степень совпадения указывает на более высокую степень сходства с эталонным

переводом и более высокий балл. Внятность и грамматика не учитываются.

Принцип оценки Bleu основан на N-граммах и штрафном факторе. N-грамма – это статическая языковая модель, которая может представлять предложение как последовательность n последовательных слов и использовать информацию о сопоставлении соседних слов в контексте для вычисления вероятности предложения, тем самым оценивая, является ли предложение соответствующим.

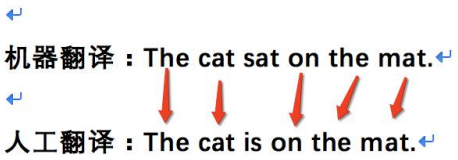

Рис.2 Пример для 1-грамма

Перевод состоит из 6 слов, и 5 слов соответствуют, поэтому степень соответствия составляет 5/6.

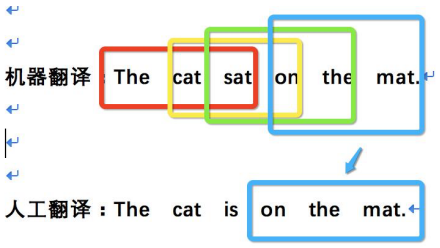

Рис.3 Пример для 3-грамма

Штрафной фактор компенсирует сгенерированные переводы, которые являются слишком короткими по сравнению с ближайшей опорной длиной с экспоненциальным затуханием.

Совпадение n-граммов подсчитывает, сколько униграмм, биграмм, триграмм и четырехграмм (i=1,...,4) соответствуют их n-граммовому аналогу в ссылочных переводах. Этот термин действует как метрика точности. Униграммы учитывают адекватность, в то время как более длинные n-граммы учитывают беглость перевода. Чтобы избежать перерасчета, количество n-граммов обрезается до максимального количества n-граммов, встречающегося в ссылке ().

В соответствии с источником [4] формула расчета оценки Bleu:

$$\text{BLEU} = \underbrace{\min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right)}_{\text{brevity penalty}} \underbrace{\left(\prod_{i=1}^{4} precision_i\right)^{1/4}}_{\text{n-gram overlap}}$$

где:

$$precision_i = \frac{\sum_{\text{snt} \in \text{Cand-Corpus}} \sum_{i \in \text{snt}} \min(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{\text{snt'} \in \text{Cand-Corpus}} \sum_{i' \in \text{snt'}} m^{i'}_{cand}}$$

Где

- m_{cand}^{i} количество i-граммов в переводе-кандидате, соответствующих эталонному переводу.
- m_{ref}^{i} количество i-граммов в эталонном переводе.
- w_{t}^{i} количество i-граммов в переводе кандидате.

Формула состоит из двух частей: brevity penalty(штрафной фактор) and the n-gram overlap (n-грамм совпадения).

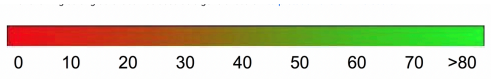

Рис.4 Шкала оценки bleu

Выводы о bleu:

1.Не учитывает точность языкового выражения (грамматики);
2.Точности оценки будут мешать общие слова;
3.Точность коротких предложений перевода иногда выше;
4.Несоблюдение синонимов или подобных выражений может привести к отклонению разумного перевода;
5. Быстрое, несложное решение для оценки перевода

Анализ:
Для анализа было выбрано 3 текста на английском языке с эталонным переводом: простой текст о семье, технический и экономический. Далее эти тексты были переведены в Google, Яндекс и Systran переводчиках. Все переводы были загружены для анализа в формате текстовых файлов.

После этого была написана программа на языке Python, выполняющая вычисление word error rate и bleu score с помощью библиотек nltk и jiwer соответственно. Кроме того было посчитана характеристика «Quality, показывающая в процентном отношении качество перевода на основе характеристик WER и Bleu Score.
Quality = (((1-WER)+Bleu_score)/2)*100%
После были построены диаграммы, визуализирующие результаты анализа.

Результаты анализа:

Результат для простого текста:

Переводчик	WER	BLEU Score	Quality
Google	0.18490566037735848	0.6682462974254928	74%
Yandex	0.18490566037735848	0.7115352865235892	76%
Systran	0.22641509433962265	0.6148330404963859	69%

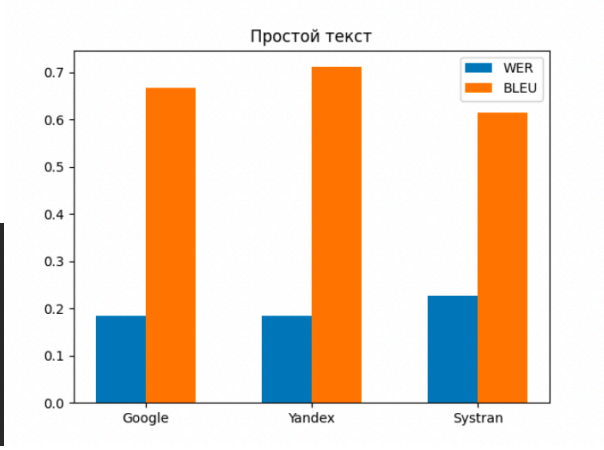

Рис.5 Результаты анализа перевода простого текста

По результатам анализа перевода простого текста видно, что худшие результаты показал Systran. Также хорошо просматривается обратная корреляция у WER и Bleu score. Лучший результат перевода у Яндекс переводчика, немного хуже у Google.

Результат для технического текста:

Переводчик	WER	BLEU Score	Quality
Google	0.6910569105691057	0.16998870353847406	23%
Yandex	0.6097560975609756	0.2305729959632378	31%
Systran	0.8048780487804879	0.05837412485999826	12%

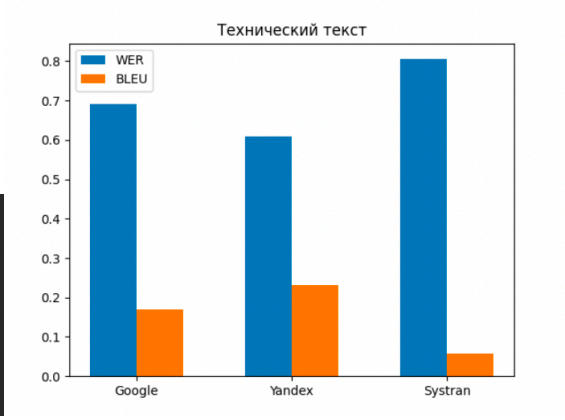

Рис.6 Результаты анализа перевода технического текст

С техническим текстом переводчики справились гораздо хуже, чем с простым. По результатам анализа перевода технического текста видно, что худшие результаты, как и с простым текстом, показал Systran. По-прежнему хорошо просматривается обратная корреляция у WER и Bleu score. Лучший результат перевода снова у Яндекс переводчика и средний результат у Google.

Результат для экономического текста:

Переводчик	WER	BLEU Score	Quality
Google	0.6883241758241758	0.20567523589393313	25%
Yandex	0.7060439560439561	0.12436471819726339	20%
Systran	0.7379273504273505	0.16777771131989447	21%

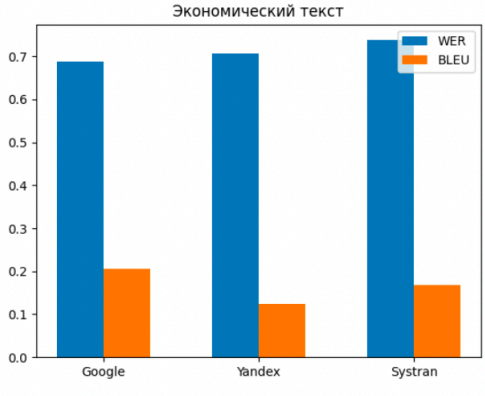

Рис.7 Результаты анализа перевода экономического текста

С экономическим текстом переводчики справились также плохо, как и с техническим. По результатам анализа перевода видно, что худшие результаты на этот раз у Yandex. По-прежнему хорошо просматривается обратная корреляция у WER и Bleu score. Лучший результат на этот раз показал Google переводчик и средний результат у Systran.

Вывод

В ходе выполнения научной исследовательской работы были разобраны такие характеристики анализа перевода, как word error rate и bleu score. Было выяснено, что они имеют обратную корреляцию. Также было проведено сравнение трех лучших переводчиков, использующих технологию машинного обучения. Было выяснено, что на настоящий момент подобные сервисы прекрасно справляются с переводом простых текстов, но для узких тематических не подходят.

Список литературы

1. BLEU. [Электронный ресурс]. – URL: https://en.wikipedia.org/wiki/BLEU (дата обращения: 20.04.2022).
2. Распознавание речи, word error rate. [Электронный ресурс]. – URL: https://docs.microsoft.com/ru-ru/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data (дата обращения: 20.04.2022).
3. Foundation of NLP Explained-Bleu Score and wer metrics. [Электронный ресурс]. – URL: https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b (дата обращения: 19.04.2022).
4. Evaluating models [Электронный ресурс]. – URL: https://cloud.google.com/translate/automl/docs/evaluate (дата обращения: 17.04.2022)
5. Bleu Score in Python. [Электронный ресурс]. – URL: https://www.journaldev.com/46659/bleu-score-in-python (дата обращения: 18.04.2022).