ВЛИЯНИЕ РАЗЛИЧНЫХ ТИПОВ ОРФОГРАФИЧЕСКИХ ОШИБОК НА КАЧЕСТВО СТАТИСТИЧЕСКОГО МАШИННОГО ПЕРЕВОДА

Мещерякова Е. М. (mescheryakova@yandex-team.ru),
Галинская И. Е. (galinskaya@yandex-team.ru),
Гусев В. Ю. (vgoussev@yandex-team.ru),
Шматова М. С. (mashashma@yandex-team.ru)

Яндекс, Москва, Россия

В статье рассматривается рост качества машинного перевода в зависимости от исправления различных типов ошибок в исходном тексте на материале трех языковых пар (англо-русской, немецко-русской и польско-русской). Мы выбрали по 500 случайных пользовательских запросов к сервису машинного перевода, последовательно исправили в них разные типы опечаток и ошибок: отсутствующую диакритику; опечатки любого рода; неправильную пунктуацию и капитализацию; все ошибки. Всего для немецкого и польского языков мы получили по пять тестовых наборов (включая оригинал), для английского — четыре (в нем отсутствует диакритика). Все наборы были протестированы на трех бесплатных статистических системах машинного перевода, и для каждого было измерено значение BLEU.

Исправление всех опечаток дает увеличение BLEU примерно на 10–15% по сравнению с оригинальными запросами. Исправление опечаток и ошибок в пунктуации и капитализации по отдельности дают улучшение примерно на 5–10% в зависимости от языка и особенностей тестового набора. Исправление же только диакритики прироста почти не дает: 0% для немецкого языка и 0,5–1% для польского.

Ключевые слова: статистический машинный перевод, качество машинного перевода, метрика BLEU, опечатки, капитализация, пунктуация

IMPACT OF DIFFERENT TYPES OF SPELLING MISTAKES ON THE QUALITY OF STATISTICAL MACHINE TRANSLATION

Mescheryakova E. M. (mescheryakova@yandex-team.ru),
Galinskaya I. E. (galinskaya@yandex-team.ru),
Gusev V. Yu. (vgoussev@yandex-team.ru),
Shmatova M. S. (mashashma@yandex-team.ru)

Yandex, Moscow, Russia

Errors in the original text will most probably affect the quality of machine translation. It would be interesting to see how different types of errors can influence the translation. To do this, we selected three sets of 500 random queries in English, German and Polish. In each set we corrected different types of errors: 1) missing diacritical marks (except English); 2) all misprints (including diacritics); 3) errors in punctuation and use of capitals; 4) all types of errors listed in 1)–3). As a result we had five sets of 500 queries for German and Polish and four sets for English.

Then we translated all the sets into Russian using three free online statistical machine translation systems and compared their BLEU scores to see how they increase in corrected tests as compared to the original ones.

We also used different types of BLEU: along with the usual one, which treats punctuation signs as words, we used simplified BLEU which disregards punctuation, and also extended BLEU which takes into consideration both punctuation and use of capitals.

We show that in a fully corrected text BLEU increases by approx. 10–15% as compared to original sets. Correcting each of the two main types of errors — misprints and punctuation/capitalization — gives an increase of 5–10% each depending on the language and on the peculiarities of the test sets. On the other hand, correcting only diacritics has very small impact on the translation quality: close to zero in German and 0,5–1% in Polish.

Key words: statistical machine translation, machine translation quality, BLEU, misprints, capitalization, punctuation

1. Введение

Влияние грамотности исходного текста на качество машинного перевода кажется очевидным: орфографические ошибки[1] в оригинале затрудняют распознавание словоформ и их последовательностей и, соответственно, процесс перевода (см., например, Carrera et al. 2009; Plesco, Rychtyckyj 2012). Отсутствие опечаток/орфографических ошибок, правильная расстановка знаков препинания и заглавных букв входит — наряду со стилистическими требованиями — в число условий подготовки текста для машинного перевода (ср. понятие «controlled/standard language»; см. Aikawa et al. 2007 о сравнительном влиянии нескольких стилистических и орфографических факторов на итоговое качество перевода). Соответственно, устранение опечаток может быть частью предобработки текста, предназначенного для любого автоматического анализа (см., например, Shoukry, Rafea 2012 о нормализации арабских текстов из твиттера для сентимент-анализа).

С распространением бесплатных онлайн-сервисов машинного перевода все большую долю в них занимают тексты, которые не подвергаются предварительному редактированию, могут содержать большое количество опечаток, неверную пунктуацию и употребление заглавных букв (Jiang et al. 2012). Добавим сюда случаи перевода интернет-страниц, элементы которых, будучи скопированы в окно перевода, превращаются в набор слов и словосочетаний, не разделенных пунктуацией.

Наконец, отдельную проблему представляет употребление диакритики в тех языках, где она существует: зачастую авторы разного рода «субстандартных» текстов склонны пренебрегать ею. Обычно это происходит при наборе текста на клавиатуре, предназначенной для другого языка и не имеющей нужных символов; таким образом, в отличие от других типов ошибок, возникающих по вине пользователя (по небрежности или неграмотности), диакритика чаще отсутствует по техническим причинам.

Некоторые онлайн-системы перевода помогают пользователю исправить опечатки в исходном тексте — указывая на слово, содержащее ошибку, или даже предлагая варианты исправления. Однако при переводе с иностранного языка воспользоваться такими подсказками, очевидно, сложно: не зная иностранного слова даже на уровне понимания общего смысла, пользователь в большинстве случаев не будет уверен и в его написании. Можно предположить, что обработка исходного текста должна производиться автоматически, являясь составной частью алгоритма машинного перевода.

Интересным кажется узнать, насколько различные типы ошибок в запросе влияют на итоговое качество перевода. Для нашего исследования мы взяли три языковые пары: англо-русскую, немецко-русскую и польско-русскую. Для каждой пары был взят набор из 500 случайных пользовательских запросов к сервису машинного перевода, которые последовательно тестировались:

---

[1] Для наших целей неважно различие между орфографическими и грамматическими ошибками, происходящими от недостаточного знания автором правил данного языка, и случайными опечатками.

- в оригинальном виде, как они задавались пользователями;
- только с исправленной диакритикой (кроме английского);
- с исправленной диакритикой и прочими опечатками/ошибками, включая лишние или недостающие пробелы;
- с исправленной капитализацией и пунктуацией;
- полностью исправленными.

В разделе 2 мы приведем примеры типичных ошибок в пользовательских запросах; в разделе 3 более подробно охарактеризуем тестовые наборы; в разделе 4 опишем последовательность эксперимента и применявшиеся в нем метрики. Наконец, в разделе 5 будут приведены собственно результаты измерений и комментарии к ним.

2. Типичные ошибки в запросах

Приведем типичные примеры ошибок в запросах (все примеры сконструированы нами либо взяты из открытых интернет-источников).

1. Опечатки:

- случайные опечатки, пропуск или перестановка одной или нескольких букв, вставка или пропуск пробела и т. д.: англ. *chanel* → *channel* ‘канал’, *theey* → *they* ‘они’, *sayi ng* → *saying* ‘говоря’; *dont’* → *don’t*; польск. *Warszwa* → *Warszawa* ‘Варшава’;
- отсутствие пробела между словами, разделенными знаками препинания: англ. *I saw him yesterday.He said…* → *I saw him yesterday. He said…* ‘Я видел его вчера. Он сказал…’. Такого рода ошибки часто бывают систематическими (т. е. автор последовательно не ставит пробелы после знаков препинания), и, поскольку многие системы не умеют разделять слова без пробела, приводят к резкому снижению качества перевода.

2. Отсутствие диакритики:

- польск. *zolte zloto* → *żółte złoto* ‘желтое золото’, нем. *mude* → *müde* ‘усталый’ (автор пользуется клавиатурой, не имеющей клавиш для особых символов и диакритических знаков). Пропуск диакритики на отдельных буквах (не систематически во всем запросе), по крайней мере, в рассматриваемых здесь языках встречается редко, потому что символы с диакритикой расположены на отдельных клавишах и обычно не по соседству с соответствующими простыми символами; единственное исключение — *l* и *ł* на соседних клавишах в польской раскладке клавиатуры. Разумеется, неиспользование диакритики возможно только при неформальной переписке.

3. Отсутствие капитализации и/или пунктуации:

- нем. *sind diese probleme für dich so wichtig* → *Sind diese Probleme für dich so wichtig?* ‘Эти проблемы для тебя так важны?’ (пользователь в переписке или в чате пренебрегает заглавными буквами и знаками препинания; как и в случае с диакритикой, такого рода «послабления» возможны только в неформальных текстах);

- польск. *data i miejsce urodzenia adres nr dokumentu tożsamości/paszportu* → *Data i miejsce urodzenia, adres, nr dokumentu tożsamości/paszportu* ‘Дата и место рождения, адрес, номер документа, удостоверяющего личность / паспорта’ (при копировании содержимого веб-страницы названия полей, которые требуется заполнить, сливаются в единую последовательность слов);
- польск. *Mam na sprzedaż nowy VW. AUTO W STANU BARDZO DOBRYM* ‘Продаю новый VW. Машина в очень хорошем состоянии’ (автор выделяет заглавными буквами важную часть сообщения);
- особый случай представляют собой обращения и приветствия в письмах, которые часто отделяются запятой, а далее идет текст с заглавной буквы и с новой строки; при копировании в окно перевода разрыв строк исчезает, и запятая оказывается перед заглавной буквой: польск. *Witam, Uprzejmie informuję…* ‘Добрый день. С уважением сообщаю…’ (обычная формулировка для официального письма).

Отметим, что в одном запросе могут содержаться и часто содержатся разные типы ошибок.

3. Характеристика тестовых наборов

Каждый из трех тестовых наборов, использованных в эксперименте, состоит из 500 случайных запросов к системе машинного перевода, каждый — длиной не более 1000 символов. Свойства наборов, однако, довольно сильно отличаются, что вызвано, с одной стороны, свойствами языков, с другой — разницей в тематике запросов.

В каждом языке могут быть свои особенности — в том числе орфографические, — которые могут приводить к ошибкам в запросах. Так, немецкий и польский языки различаются по количеству букв с диакритикой: 3 в немецком и 9 в польском; соответственно, игнорирование диакритики в польском языке сильнее искажает текст, и можно предположить, что и степень влияния этого фактора на качество перевода будет выше. С другой стороны, в немецком языке принято писать с большой буквы все имена существительные; соответственно, пренебрежение капитализацией вызывает дополнительные орфографические ошибки.

Тематика рассматриваемых тестовых наборов также достаточно сильно отличается. Так, к примеру, в переводах с немецкого языка около 40 % составляют учебные тексты и упражнения, еще около 25 % — литературные тексты, переводы веб-страниц — 9 %, а переписка любого рода (включая чаты) — лишь 8 %. Напротив, в запросах на перевод с английского языка доля переписки превышает 30 %, доля учебных текстов составляет около 20 %, примерно столько же составляют переводы веб-страниц, а доля литературных текстов — всего около 8 %. В польском же переводы веб-страниц находятся в лидерах — почти 45 %; следом идет переписка (около 40 %), а литературные и учебные тексты вместе составляют около 10 %.

Очевидно, разница в тематике является одной из причин различной средней длины запросов, см. Таблицу 1.

Таблица 1. Средняя длина запроса

Язык	Среднее количество слов в запросе
Английский	17
Немецкий	23
Польский	20

Можно было бы ожидать, что такое различие в тематике скажется на среднем уровне грамотности запросов: литературные тексты (которые, скорее всего, не набираются вручную, а копируются из какого-либо источника в интернете) должны быть орфографически и пунктуационно выверены, в отличие, к примеру, от сообщений в чатах. В этом случае средний уровень грамотности немецких запросов должен быть выше, чем, скажем, английских. Сравним, однако, данные о количестве запросов с ошибками разных типов в Таблице 2:

Таблица 2. Доля запросов с разными типами ошибок в тестовых наборах

Язык	Диакритика	Опечатки (включая диакритику)	Капитализация + пунктуация	Ошибки любого типа
Английский	—	32,4%	38%	53,2%
Немецкий	5%	40,2%	48,8%	67,2%
Польский	12%	36,6%	62,2%	71,4%

Мы видим, что наши ожидания не полностью оправдываются. В английском тестовом наборе, несмотря на большое количество потенциально «ненормативной» переписки, доля ошибочных запросов как в целом, так и по отдельным типам ошибок ниже всего. Количество ошибок в употреблении заглавных букв и в пунктуации в польском языке значительно превышает количество аналогичных ошибок в немецком.

Часть этих различий, тем не менее, можно объяснить. Соотношение ошибок в диакритике в польском и немецком языке в целом соответствует ожиданиям. Большое количество ошибок на капитализацию и пунктуацию в польском языке, очевидно, происходит из-за переводов веб-страниц, отдельные элементы которых при копировании сливаются (см. примеры ошибок выше).

4. Методика эксперимента

Для сравнения результатов в каждом из трех тестовых наборов последовательно исправлялись ошибки различных типов. В результате, для каждого языка, помимо оригинального, были созданы следующие четыре набора:

- с исправленной диакритикой (кроме английского);
- с исправленной диакритикой и прочими опечатками/ошибками, включая лишние или недостающие пробелы;

- с исправленными капитализацией и пунктуацией;
- полностью исправленные.

Всего, таким образом, для немецкого и польского языков у нас было по пять наборов по 500 запросов, для английского — четыре.

Для каждого языка были подготовлены эталоны переводов на русский язык, после чего все наборы были протестированы на трех бесплатных статистических онлайн-системах машинного перевода. Для оценки качества применялась метрика BLEU (Bilingual Evaluation Understudy), широко используемая для оценки статистического машинного перевода.

BLEU основана на сравнении машинного перевода с эталоном, сделанным человеком. Для этого подсчитывается количество последовательностей из n слов (n-граммов), совпадающих в сравниваемом переводе и в эталоне; n обычно берется от 1 до 4. Значение BLEU высчитывается в среднем для всего корпуса переводов (в нашем случае 500 фрагментов для каждого языка) и составляет от 0 до 1 (либо от 0 до 100), где 0 означает, что совпадения отсутствуют, а 1 (100) — что сравниваемый корпус полностью идентичен эталону (см. подробнее о метрике BLEU: Papineni et al. 2002).

В нашем эксперименте использованы три метрики BLEU: а) стандартная (учитывающая пунктуационные знаки как отдельные токены); б) BLEU без учета пунктуации; и в) BLEU с учетом пунктуации и капитализации (т. е. учитывающая также различие строчных и заглавных букв в переводе и эталоне).

5. Результаты эксперимента

В этом разделе приводятся данные по изменению BLEU в трех системах машинного перевода для каждого языкового набора при исправлении ошибок разного типа: опечаток, капитализации и пунктуации, всех ошибок. В скобках указывается прирост значения BLEU по сравнению с исходными запросами.

5.1. Приведем результаты подсчетов стандартного BLEU (с учетом пунктуации)

Таблица 3

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Английский язык				
С 1	28,8	30,7 (+1,9)	30,1 (+1,3)	32,1 (+3,3)
С 2	30,9	33,1 (+2,2)	32,1 (+1,2)	34,5 (+3,6)
С 3	26,6	28,0 (+1,4)	28,9 (+2,3)	30,2 (+3,6)

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Немецкий язык				
C 1	23,9	26,2 (+2,3)	24,2 (+0,3)	26,9 (+3,0)
C 2	22,6	24,4 (+1,8)	23,0 (+0,4)	25,4 (+2,8)
C 3	20,4	21,8 (+1,4)	20,8 (+0,4)	22,2 (+1,8)
Польский язык				
C 1	33,1	35,0 (+1,9)	37,7 (+4,6)	40,0 (+6,9)
C 2	20,9	22,0 (+1,1)	26,1 (+5,2)	27,3 (+6,4)
C 3	20,0	20,6 (+0,6)	24,2 (+4,2)	24,9 (+4,9)

Прокомментируем результаты измерений.

1. В английском языке исправление обоих типов ошибок дает более или менее равномерный прирост качества (возможно, это соотносится с тем, что процент запросов с ошибками каждого из этих типов в английском сравним — 32,4% и 38%, см. Таблицу 1). Правда, в разных системах вклад двух типов ошибок может отличаться: в C1 и C2 сильнее влияние опечаток, в C3 больше влияет капитализация/пунктуация.
2. В немецком обращает на себя внимание существенно большее влияние на рост BLEU исправление опечаток по сравнению с исправлением ошибок капитализации и пунктуации — хотя процент запросов с ошибками второго типа не меньше, а даже несколько больше, чем первого (40,2% и 48,8% соответственно — соотношение близко к тому, которое мы видели в английском наборе). Сходство соотношения во всех системах подтверждает этот результат.
3. В польском языке, напротив, очень сильно влияние капитализации и пунктуации и невелик вклад опечаток. В определенной степени это связано с особенностями тестового набора и большой долей таких ошибок в нем (62,2% запросов), однако это, очевидно, не единственная причина: по сравнению с английским доля запросов с опечатками в польском несколько выше (36,6% против 32,4%), а влияние их исправления на качество ниже во всех системах перевода.

Посмотрим, насколько влияет на качество перевода исправление только диакритики, без прочих опечаток, в польском и немецком языках. Приводятся только значения BLEU при исправленной диакритике и разница с исходными запросами.

Таблица 4

	Немецкий	Польский
C 1	23,8 (−0,1)	33,8 (+0,7)
C 2	22,6 (+0,0)	21,0 (+0,1)
C 3	20,4 (+0,0)	20,1 (+0,1)

Как видно, исправление только диакритики дает очень небольшой эффект. В немецком языке он вовсе нулевой. В польском он отличен от нуля — что, видимо, объясняется бóльшим количеством букв с диакритикой и, соответственно, бóльшим процентом ошибок данного типа (см. Таблицу 2), — но тоже очень невелик. Более или менее заметен он в польском языке только в системе 1, даже несмотря на то, что и общие показатели BLEU для польского языка в ней выше; в процентном отношении прирост BLEU в Системе 1 составляет 2,1%, в Системах 2 и 3 — 0,5%.

Приведем для сравнения результаты измерений по другим метрикам BLEU.

5.2. BLEU без учета пунктуации

Таблица 5

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Английский язык				
C 1	24,4	26,3 (+1,9)	24,9 (+0,5)	26,7 (+2,3)
C 2	26,9	28,8 (+1,9)	26,8 (−0,1)	28,8 (+1,9)
C 3	23,2	24,6 (+1,4)	23,7 (+0,5)	25,2 (+2,0)
Немецкий язык				
C 1	18,5	20,6 (+2,1)	18,7 (+0,2)	21,2 (+2,7)
C 2	17,5	19,1 (+1,6)	17,5 (+0,0)	19,7 (+2,2)
C 3	16,1	17,1 (+1,0)	16,1 (+0,0)	17,3 (+1,2)
Польский язык				
C 1	29,7	31,8 (+2,1)	30,0 (+0,3)	32,2 (+2,5)
C 2	17,1	18,2 (+1,1)	18,4 (+1,3)	19,4 (+2,3)
C 3	16,2	16,9 (+0,7)	16,8 (+0,6)	17,3 (+1,1)

Ожидаемым образом, здесь резко уменьшаются цифры в предпоследней колонке, в немецком языке — вовсе до нуля. Оставшиеся цифры, очевидно, показывают реальный вклад исправления капитализации и пунктуации в качество перевода как такового. В то же время влияние исправленных опечаток в немецком и польском языках при таком способе измерения несколько повышается.

5.3. BLEU с учетом капитализации и пунктуации

Таблица 6

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Английский язык				
C 1	26,7	28,4 (+1,7)	28,8 (+2,1)	30,6 (+3,9)
C 2	29,3	31,4 (+2,1)	30,8 (+1,5)	33,2 (+3,9)
C 3	24,9	26,2 (+1,3)	27,6 (+2,7)	28,8 (+3,9)
Немецкий язык				
C 1	21,7	24,0 (+2,3)	22,4 (+0,7)	24,9 (+3,2)
C 2	21,3	23,0 (+1,7)	21,9 (+0,6)	24,1 (+2,8)
C 3	19,1	20,4 (+1,3)	19,6 (+0,5)	21,0 (+1,9)
Польский язык				
C 1	30,1	31,7 (+1,6)	36,5 (+6,4)	38,7 (+8,6)
C 2	19,5	20,5 (+1,0)	25,0 (+5,5)	26,3 (+6,8)
C 3	18,3	18,9 (+0,6)	23,1 (+4,8)	23,9 (+5,6)

Заключение

Данная работа посвящена влиянию орфографической правильности исходного текста на качество статистического машинного перевода. Мы показали, что исправление всех — орфографических и пунктуационных — ошибок может дать прирост качества по метрике BLEU примерно на 10–15 % по сравнению с неисправленным текстом. Что касается отдельных классов ошибок: орфографических, с одной стороны, и в пунктуации и капитализации — с другой, вклад каждого из этих классов может различаться в зависимости от свойств языка и особенностей запросов. Мы также рассмотрели влияние на перевод отдельного типа ошибок — отсутствия диакритики, которое может возникать не только по небрежности пользователя, но и по техническим причинам. Вклад исправления этого типа ошибок в качество перевода оказался невелик — впрочем, он в большей мере, чем другие, зависит от конкретного языка.

Было бы интересно проверить полученные результаты на других языковых парах. Кроме того, в дальнейшем необходимо исследовать влияние на качество перевода других типов ошибок — в первую очередь синтаксических.

Литература

1. *Aikawa T., Schwartz L., King R., Corston-Oliver M., Lozano C.* Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. MT Summit XI, 10–14 September. Proceedings. Copenhagen, 2007, pp. 1–7.
2. *Shoukry A., Rafea A.* Preprocessing Egyptian dialect tweets for sentiment mining. AMTA-2012: Fourth workshop on computational approaches to Arabic script-based languages. Proceedings. San Diego, 2012, pp. 47–56.
3. *Jiang J., Way A., Haque R.* Translating user-generated content in the social networking space. AMTA-2012: the Tenth Biennial Conference of the Association for Machine Translation in the Americas. Proceedings. San Diego, 2012.
4. *Carrera J., Beregovaya O., Yanishevsky A.* (2009). Machine Translation for Cross-Language Social Media, available at: http://www.promt.com/company/technology/pdf/machine_translation_for_cross_ language_social_media.pdf
5. *Papineni K., Roukos S., Ward T., Zhu W. J.* (2002). BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. Proceedings. Stroudsburg, 2002, pp. 311–318.
6. *Pesko C., Rychtyckyi N.* Machine Translation as a Global Enterprise at Ford. AMTA-2012: the Tenth Biennial Conference of the Association for Machine Translation in the Americas. Proceedings. San Diego, 2012.