ДОРОЖКА ПО ОЦЕНКЕ МАШИННОГО ПЕРЕВОДА ROMIP MTEVAL 2013: ОТЧЕТ ОРГАНИЗАТОРОВ

Браславский П. (pbras@yandex.ru)
Kontur Labs; Уральский федеральный университет,
Екатеринбург, Россия

Белобородов А. (xander-beloborodov@yandex.ru)
Уральский федеральный университет,
Екатеринбург, Россия

Шаров С. (s.sharoff@leeds.ac.uk)
University of Leeds, Лидс, Великобритания

Халилов М. (maxim@tauslabs.com)
TAUS Labs, Амстердам, Нидерланды

Ключевые слова: машинный перевод, оценка, англо-русский перевод

ROMIP MT EVALUATION TRACK 2013: ORGANIZERS' REPORT

Braslavski P. (pbras@yandex.ru)
Kontur labs; Ural Federal University, Russia

Beloborodov A. (xander-beloborodov@yandex.ru)
Ural Federal University, Russia

Sharoff S. (s.sharoff@leeds.ac.uk)
University of Leeds, Leeds, UK

Khalilov M. (maxim@tauslabs.com)
TAUS Labs, Amsterdam, Netherlands

The paper presents the settings and the results of the ROMIP 2013 machine translation evaluation campaign for the English-to-Russian language pair. The quality of generated translations was assessed using automatic metrics and human evaluation. We also demonstrate the usefulness of a dynamic mechanism for human evaluation based on pairwise segment comparison.

Keywords: machine translation, evaluation, English-to-Russian translation

1. Введение

Русский и английский были одной из первых языковых пар на заре исследований в этой области машинного перевода (МП) в 1950-х годах [Hutchins2000]. С тех пор парадигмы МП поменялись много раз, многие системы для этой языковой пары появлялись и исчезали, но, насколько нам известно, до сих пор не проводилась систематическая сравнительная оценки систем МП, аналогичная DARPA'94 [White et al., 1994] и более поздним мероприятиям. Семинар по статистическому машинному переводу (Workshop on Statistical Machine Translation, WMT) в 2013 году впервые включил русско-английскую пару в свою программу.[1] На данный момент эта оценка еще не проведена, к тому же в семинаре примут участие системы, обученные на данных, предоставленных организаторами. За рамками оценки останутся существующие системы, в частности — системы на основе правил и гибридные системы.

Кампании по оценке играют важную роль в развитии технологий МП. В последнее время был проведен ряд открытых кампаний для различных комбинаций европейских, азиатских и семитских языков, см. [Callison-Burch et al., 2011; Callison-Burch et al., 2012; Federico et al., 2012]. В этой статье мы описываем кампанию по оценке англо-русского машинного перевода в рамках РОМИП.

РОМИП (Российский семинар по Оценке Методов Информационного Поиска)[2] — это российский аналог TREC и других инициатив по оценке задач информационного поиска. Первый цикл оценки был организован в 2002 году. В течение этих десяти лет РОМИП организовал серию дорожек по оценке, включая классическую задачу поиска по запросу, задачи тематической классификации документов, вопросно-ответного поиска, формирования сниппетов, анализа тональности текста, поиска изображений и т. д. В рамках этой деятельности было подготовлено несколько свободно распространяемых наборов данных, содержащих документы и оценки релевантности, сделанные асессорами. Российские сообщества, занимающиеся информационным поиском и машинным переводом, имеют давние связи, их представители тесно общаются. Поэтому было естественным организовать кампанию по оценке МП в рамках РОМИП, используя накопленный опыт семинара. Кроме того, важной целью мероприятия была консолидация групп, разрабатывающих как статистические системы МП (SMT), так и системы, основанные на правилах (RBMT).

Одна из проблем для систем МП, работающих с русским языком, и для их оценки — это необходимость иметь дело с относительно свободным порядком слов в предложении и развитой морфологией. За счет развитой морфологии у русских лемм много словоформ (в среднем 8,2 формы для существительных, 34,6 — для глаголов [Sharoff et al., 2013]), что осложняет выравнивание на уровне слов при статистическом подходе. Дистантные зависимости создают дополнительные проблемы, особенно для SMT-систем.

---

[1] http://www.statmt.org/wmt13/

[2] http://romip.ru

Для оценки было выбрано одно направление перевода (английский → русский). Во-первых, для этого направления нам намного проще было найти асессоров, для которых целевой язык является родным. Во-вторых, системы-участницы в основном используются именно в этом направлении (перевод английских текстов для русскоязычных пользователей).

2. Данные

При формировании тестового корпуса текстов мы руководствовались двумя соображениями. Во-первых, известно, что предметная область и жанр текста влияют на качество перевода [Langlais, 2002; Babych et al., 2007]. Таким образом, мы хотели обеспечить хотя бы минимальное жанровое разнообразие текстов, входящих в корпус. Во-вторых, мы хотели использовать источники, допускающие дальнейшее распространение текстов по лицензии Creative Commons. В итоге корпус был сформирован из двух источников, соответственно — из текстов двух жанров. Новостные тексты было собраны с английского раздела Wikinews[3]. Формальные тексты (регламенты, инструкции, положения, официальные документы) был собраны из Веба с использованием жанрового классификатора [Sharoff, 2010]. После применения автоматической классификации был проведен ручной отбор текстов.

Начальный корпус состоял из 8356 оригинальных документов общим объемом 148 864 английских предложений. В корпусе были представлены оригинальные документы целиком, т. к. некоторые системы могут использовать для перевода контекст предложения. Источник 100 889 предложений в корпусе — Wikinews; 47 975 предложений относятся к формальным текстам. Первые 1002 предложения были опубликованы заранее, чтобы участники могли адаптировать свои системы к используемому формату. Так как корпус был подготовлен полностью автоматически, он не лишен дефектов (например, часто встречается некорректная разбивка на предложения, остатки HTML-разметки и т. п.). Участники должны были прислать организаторам русские переводы 147 862 предложений в течение недели после публикации исходного тестового корпуса.

Примеры предложений тестового корпуса:

90237	*Ambassadors from the United States of America, Australia and Britain have all met with Fijian military officers to seek insurances that there wasn't going to be a coup.*
102835	*If you are given a discount for booking more than one person onto the same date and you later wish to transfer some of the delegates to another event, the fees will be recalculated and you will be asked to pay additional fees due as well as any administrative charge.*

[3] http://en.wikinews.org/

Тексты в исходном корпусе не были до этого переведены на русский язык, т. е. системы-участники не могли заранее использовать переводы для обучения. Для оценки мы выбрали 947 «чистых» предложений (т. е. с корректными границами, без паразитной HTML разметки и т. п.), из них 759 — новостных и 188 — из формальных текстов.

Эти предложения примерно равными порциями были назначены для перевода трем переводчикам (переводчик 1: предложения 1–316; переводчик 2: 317–632; переводчик 3: 633–947). Переводчики 1 и 2 сообщили, что они потратили от 20 до 30 часов на перевод всего задания. Оба переводчика сообщили, что время, потраченное на перевод отдельного предложения значительно различалось. В отличие от перевода связного текста, дополнительная сложность возникает из-за необходимости переключаться между темами и понимать контекст предложения (переводчикам иногда приходилось обращаться к набору данных, содержащему оригинальные документы). Переводчик 3 не смог выполнить перевод в срок, поэтому ему принадлежат только 152 перевода в третьей порции. Остальные предложения переведены двумя членами одной из групп-участниц. Все 947 переводов использовались для автоматической оценки качества переводов, 330 предложений из 947 были выбраны для ручной оценки (190 новостных и 140 формальных текстов).

Дополнительно мы сделали объявление в списке рассылки конкурса, нескольких онлайн форумах переводчиков и в группах Facebook с просьбой принять участие в коллективном переводе тестовых предложений на сайте TranslatedBy.[4] Сравнение профессионального и коллективного перевода — тема отдельного исследования.

Дополнительно организаторы предоставили участникам доступ к следующим ресурсам:

- 1M предложений англо-русского параллельного корпуса, распространяемого Яндексом (этот корпус используется в WMT13)[5];
- 119K предложений англо-русского параллельного корпуса из репозитория TAUS.

Эти наборы данных не связаны с корпусом, который был подготовлен в рамках кампании по оценке; цель этих дополнительных данных — снизить порог участия для групп, которые не имеют собственных данных достаточного объема для этого направления перевода.

3. Ручная и автоматическая оценка

Основной принцип, который мы хотели реализовать в ручной оценке, — сделать оценку как можно более простой для асессора, а ее результаты — интерпретируемыми. Мы выбрали вариант ранжирования систем на основе попарных сравнений вариантов перевода. Такой подход отличается от *ранжирования* нескольких

[4] http://translatedby.com

[5] http://translate.yandex.ru/corpus

вариантов перевода асессором — подхода, который используется в рамках экспериментов по оценке WMT. В случае большого количества участвующих систем асессоры каждый раз ранжируют только часть вариантов переводов. На основе частичных рангов не всегда просто получить однозначное полное ранжирование систем [Callison-Burch et al., 2012]. На основании попарных сравнений проще построить общее ранжирование, к тому же попарные сравнения — более простая задача для асессора. Однако такой метод подразумевает больший объем оценки (который все же остается приемлемым в случае небольшого количества участвующих систем). Ниже мы обсуждаем, как можно снизить объем ручной оценки.

В нашем случае асессоры должны были делать попарные сравнения двух предложений — переводов участвующих систем — с образцовым переводом, выполненным человеком. Асессор должен был выбрать лучший из двух вариантов или отметить, что оба варианта эквивалентны. При этом асессор не видел исходное предложение, а только человеческий перевод.

Как было сказано выше, 330 тестовых предложений были задействованы в ручной оценке. Исходная идея состояла в том, чтобы генерировать пары предложений для оценки динамически для оптимизации объема оценки. К сожалению, ограничения используемого инструмента оценки не позволили реализовать такой сценарий. Мы были вынуждены проводить полное сравнение — 28 пар на одно тестовое предложение (для 8 систем, участвовавших в ручной оценке). Изначально задачи по оценке были распределены между 11 асессорами (добровольцами и членами участвующих в кампании команд) с небольшим перекрытием. Задания по оценке были распределены таким образом, чтобы все варианты перевода одного предложения оценивались одним асессором, что предположительно должно приводить к более согласованному ранжированию. Недостаток такого подхода — в том, что члены участвующих команд оценивают, в том числе, результаты работы «своих» систем. Незадолго до срока окончания оценки некоторые асессоры сообщили, что не смогут закончить оценку вовремя. Невыполненные задания были переназначены другим асессорам; дополнительно три новых асессора присоединились к оценке. Таким образом всего в оценке приняло участие 14 человек. Переводы 60 тестовых предложений были оценены с двойным перекрытием (таким образом, для 60×28=1680 пар у нас есть решение двух асессоров). Общий объем оценки составил 10 920 попарных сравнений. По сообщениям асессоров, на оценку одной пары уходило от 30 до 90 секунд, при этом для оценки некоторых сложных предложений требовалось до 5 минут.

Для оценки мы использовали многофункциональный инструмент оценки машинного перевода TAUS DQF в режиме «быстрое сравнение» (*quick comparison*).[6]

На основе оценок асессоров системы можно ранжировать для каждого предложения из тестового набора. В случае равенства очков ранги усреднялись. Например, так выглядят ранги, если системы на позициях 2–4, 7–8 имеют равное количество очков:

1 3 3 3 5 6 7.5 7.5

[6] https://tauslabs.com/dynamic-quality/dqf-tools-mt

Для получения общего ранжирования систем ранги на уровне предложений усреднялись по всем предложениям.

После того, как мы получили все попарные сравнения вариантов перевода, мы смогли провести моделирование динамического формирования пар для сравнения и понять, какой объем оценки можно сэкономить с использованием такой методики. Идея состоит в том, чтобы сначала получить предварительное ранжирование систем (например, на основе автоматических метрик), а потом сортировать этот «массив предложений» с помощью алгоритма сортировки вставками (или его варианта с использованием бинарного поиска).

В дополнение к ручной оценке мы также запустили автоматическую оценку, используя следующие метрики: BLEU [Papineni et al. 2001], METEOR [Banerjee and Lavie, 2005], TER [Snover et al., 2009] и GTM [Turian et al., 2003]. BLEU и METEOR могут рассматриваться как метрики близости машинного перевода образцовому; TER и GTM демонстрируют более высокую корреляцию с объемом необходимого постредактирования [O'Brien, 2011].

4. Результаты

Мы получили результаты от пяти участников, две команды прислали по два прогона. Таким образом, в сумме у нас было семь прогонов для оценки (обозначены P1..P7 в данном отчете), см. краткое описание систем в Табл. 1. Как видно из таблицы, в кампании приняли участие как признанные группы из индустрии и академических организаций, так и молодые команды. В оценку были включены также переводы 947 тестовых предложений четырех онлайн систем (обозначены в отчете OS1..OS4). Таким образом, в автоматической оценке участвовало 11 прогонов, в ручной — восемь (четыре онлайн системы и четыре системы-участницы; в ручной оценке не участвовали прогоны P3, P6 и P7).

Таблица 1. Участники ROMIP MTEval 2013

ID	Краткое описание системы
P1	**Compreno (ABBYY)** http://www.abbyy.ru/science/technologies/business/compreno/
P2	**Pharaon (анонимный участник)** Система на основе Moses SMT, использованы корпуса Яндекса и TAUS.
P3,4	**Balagur (Школа анализа данных)** Система на базе MOSES, использован корпус Яндекса (1M) и новостной корпус (200K), собранный по новостным сайтам.
P5	**ЭТАП-3 (ИППИ РАН)** Система перевода на основе правил, использует составленный вручную словарь примерно со 100 000 входов [Boguslavsky1995]
P6,7	**Pereved (МФТИ)** Система основана на Moses и натренирована на параллельных предложениях, извлеченных из Интернета.

Таблица 2. Результаты автоматической оценки

Метрика/ID	OS1	OS2	OS3	OS4	P1	P2	P3	P4	P5	P6	P7
Все (947 предложений)											
BLEU	0,150	0,141	0,133	0,124	0,157	0,112	0,105	0,073	0,094	0,071	0,073
METEOR	0,258	0,240	0,231	0,240	0,251	0,207	0,169	0,133	0,178	0,136	0,149
TER	0,755	0,766	0,764	0,758	0,758	0,796	0,901	0,931	0,826	0,934	0,830
GTM	0,351	0,338	0,332	0,336	0,349	0,303	0,246	0,207	0,275	0,208	0,230
Новости (759 предложений)											
BLEU	0,137	0,131	0,123	0,114	0,153	0,103	0,096	0,070	0,083	0,066	0,067
METEOR	0,241	0,224	0,214	0,222	0,242	0,192	0,156	0,127	0,161	0,126	0,136
TER	0,772	0,776	0,784	0,777	0,768	0,809	0,908	0,936	0,844	0,938	0,839
GTM	0,335	0,324	0,317	0,320	0,339	0,290	0,233	0,201	0,257	0,199	0,217

Таблица 3. Ранжирование систем на основе ручной оценки (усредненные ранги, от лучших к худшим слева направо)

Все (330 предложений)							
OS3	**P1**	**OS1**	**OS2**	**OS4**	**P5**	**P2**	**P4**
3,159	3,350	3,530	3,961	4,082	5,447	5,998	6,473
Новости (190 предложений)							
OS3	**P1**	**OS1**	**OS2**	**OS4**	**P5**	**P2**	**P4**
2,947	3,450	3,482	4,084	4,242	5,474	5,968	6,353
Формальные тексты (140 предложений)							
P1	**OS3**	**OS1**	**OS2**	**OS4**	**P5**	**P2**	**P4**
3,214	3,446	3,596	3,793	3,864	5,411	6,039	6,636
Предварительное ранжирование, сортировка вставками							
P1	**OS1**	**OS3**	**OS2**	**OS4**	**P5**	**P4**	**P2**
3,318	3,327	3,588	4,221	4,300	5,227	5,900	6,118
Предварительное ранжирование, сортировка бинарными вставками							
OS1	**P1**	**OS3**	**OS2**	**OS4**	**P5**	**P2**	**P4**
2,924	3,045	3,303	3,812	4,267	5,833	5,903	6,882

Табл. 2 содержит значения автоматических метрик для всех прогонов участников и четырех онлайн систем. По автоматическим метрикам OS1 лидирует на полном наборе тестовых предложений и на предложениях формальных документов, P1 демонстрирует лучший результат на предложениях новостных документов.

Итоговое ранжирование систем на основе ручной оценки представлено в Табл. 3. Внутри трех групп участников разница между усредненными рангами статистически незначима (по t-тесту Уэлча, уровень значимости $p=0,05$):

(OS1, OS3, P1), (OS2, OS4) и (P2, P4). Система P5 располагается между последними двумя группами. Ранжирование систем сохраняется на подмножествах тестового набора, соответствующих новостям и формальным документам. В отличие от ранжирования на основе автоматических метрик (Табл. 2) OS3 входит в тройку лидеров по результатам ручной оценки. Аналогичным образом P5 ранжируется выше, чем P2 по результатам ручной оценки, в то время как автоматические метрики располагают эти системы в обратном порядке. Это наблюдение еще раз подтверждает факт, что автоматические метрики систематически недооценивают качество систем МП, основанных на правилах [Béchar et al., 2012].

Нижняя часть Табл. 3 содержит результаты моделирования ручной оценки систем с динамическим формированием пар предложений для оценки. Системы были предварительно отсортированы на основе метрики NIST (см. Табл. 2). После этого варианты перевода для одного тестового предложения были ранжированы с помощью алгоритма сортировки вставками на основе имеющихся ручных оценок пар. В результате мы получили ранжирование систем, несколько отличающееся от ранжирования на основе полного набора оценок, т. к. при сортировке мы не использовали «усредненные» ранги.[7] При этом ранжирование можно считать идентичным — с точностью до взаимного расположения статистически различных *групп* систем. Преимущество такого подхода в том, что для ранжирования нам достаточно сделать существенно меньше попарных сравнений. В случае классической сортировки вставками нам понадобилось 5131 сравнений (15,5 на одно тестовое предложение; 56% полного набора попарных сравнений для 330 тестовых предложений и 8 систем); сортировка бинарными вставками показала себя еще лучше: 4327 сравнений (13,1 на предложение; 47% от полного набора сравнений). Предположительно, объем оценок можно снизить еще больше, если предварительно ранжировать системы на уровне отдельных предложений.

Показатели согласия асессоров аналогичны показателям при ранжировании вариантов перевода [Callison-Burch et al., 2012; Callison-Burch et al., 2011]: $\kappa=0{,}34$, $\alpha=0{,}48$. Согласованность повышается, если мы рассмотрим только сравнения трех лучших систем с остальными (т. е. не учитываем сравнения внутри групп): $\alpha=0{,}53$. Аналогично, согласованность падает, если мы учитываем только сравнения внутри группы трех лучших систем: $\kappa=0{,}23$, $\alpha=0{,}33$. Эти результаты согласуются с данными о низкой согласованности асессоров в случае оценки систем примерно одинакового уровня [Callison-Burch et al., 2011].

---

[7] Такое итоговое ранжирование не является полностью независимым от метода предварительной сортировки систем. Например, если предварительная сортировка систематически ранжирует одну систему из двух выше, а ручная оценка систематически считает их равными, то в итоговом ранжировании сохранится порядок предварительной сортировки.

5. Заключение

Это был первый опыт систематической сравнительной оценки систем машинного перевода для направления английский → русский. В будущем мы планируем построить новый тестовый корпус с более широкой жанровой палитрой. Мы постараемся дополнить оценку направлением перевода русский → английский. Мы надеемся привлечь больше участников, в том числе международных, и планируем подготовить «легкую версию» дорожки для студентов и молодых исследователей. Также мы рассмотрим проблему адаптации автоматических метрик оценки к русскоязычным данным. Такая метрика должна учитывать развитую русскую морфологию и свободный порядок слов, о чем говорилось выше. С этой целью мы планируем использовать данные ручной оценки, собранные в 2013 году.

Тестовый корпус, профессиональные переводы, переводы систем-участниц и данные ручной оценки будут доступны по адресу http://romip.ru/mteval/data/.

Благодарности

Мы хотели бы поблагодарить всех переводчиков и асессоров, а также Анну Цыганкову — за координацию проекта, Максима Губина и Марину Некрестьянову — за помощь в организации. Мы благодарны компаниям Яндекс и ABBYY, которые приняли активное участие в подготовке мероприятия и взяли на себя часть расходов, связанных с проведением оценки.

Литература

1. *Bogdan Babych, Anthony Hartley, Serge Sharoff, and Olga Mudraya.* 2007. Assisting translators in indirect lexical transfer. In Proc. of 45 ACL, pages 739–746, Prague.
2. *Satanjeev Banerjee and Alon Lavie.* 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan, June.
3. *Hanna Béchar, Raphaël Rubino, Yifan He, Yanjun Ma, and Josef van Genabith.* 2012. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In Proceedings of COLING'12, Mumbai.
4. *Igor Boguslavsky.* 1995. A bi-directional Russian-to-English machine translation system (ETAP-3). In Proceedings of the Machine Translation Summit V, Luxembourg.
5. *Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan.* 2011. Findings of the 2011 workshop on statistical machine translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 22–64. Association for Computational Linguistics.

6. *Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia.* 2012. Findings of the 2012 workshop on statistical machine translation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 10–51, Montréal, Canada, June.
7. *Marcelo Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stuker.* 2012. Overview of the iwslt 2012 evaluation campaign. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT), pages 12–34, Hong Kong, December.
8. *John Hutchins,* editor. 2000. Early years in machine translation: Memoirs and biographies of pioneers. John Benjamins, Amsterdam, Philadelphia. http://www.hutchinsweb.me.uk/EarlyYears-2000-TOC.htm.
9. *Philippe Langlais.* 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In Proceedings of Second international workshop on computational terminology (COMPUTERM 2002), pages 1–7, Taipei, Taiwan. http://acl.ldc.upenn.edu/W/W02/W02-1405.pdf.
10. *Sharon O'Brien.* 2011. Towards predicting post-editing productivity. Machine translation, 25(3):197–215.
11. *Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.* 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22 176 (W0109-022), IBM Thomas J. Watson Research Center.
12. *Serge Sharoff, Elena Umanskaya, and James Wilson.* 2013. A frequency dictionary of Russian: core vocabulary for learners. Routledge, London.
13. *Serge Sharoff.* 2010. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, Genres on the Web: Computational Models and Empirical Studies, pages 149–166. Springer, Berlin/New York.
14. *Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz.* 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 259–268, Athens, Greece, March.
15. *Joseph Turian, Luke Shen, and I. Dan Melamed.* 2003. Evaluation of machine translation and its evaluation. In Proceedings of Machine Translation Summit IX, New Orleans, LA, USA, September.
16. *John S. White, Theresa O'Connell, and Francis O'Mara.* 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and further approaches. In Proceedings of AMTA'94, pages 193–205.