Исследование методов улучшения результатов машинного перевода

Лори Тике (Lori Thicke)

Широкое распространение машинного перевода скорее говорит о том, что изменились требования к качеству перевода, нежели об улучшении технологий. Разумеется, технологии МП совершенствуются, но самое значительное достижение, возможно, заключается в том, что пользователи уже не ожидают получить «готовый» качественный перевод. Многие из них заранее вкладывают деньги в настройку программ МП, чтобы обеспечить на выходе машинный перевод, максимально близкий по качеству к выполненному специалистом.

По мере накопления опыта становится ясно, что результаты МП интересны, только когда настройку самой системы сочетают с другими видами оптимизации до, после или во время процесса машинного перевода. Независимо от того, какой подход используется в машинном переводе — основанный на правилах или статистический, — хорошо обученная система перевода, является, несомненно, самым важным компонентом. Но когда система уже оптимизирована, а итеративный процесс для ее улучшения уже внедрен, как еще можно улучшить результаты машинного перевода?

Именно этот вопрос изучает компания Lexcelera. В частности, мы хотели выяснить, какие виды оптимизации принесли бы наибольшую отдачу от капиталовложений. Поскольку совершенству нет предела, мы пытались определить, какие улучшения будут наиболее эффективными для повышения качества необработанного результата без существенного усложнения процесса. Оптимизация процесса МП чаще всего включает обучение системы терминологии на языке перевода — это также называется адаптацией. Другие способы улучшения результатов предусматривают крайне необходимый этап повторного обучения системы на материалах реальных проектов, а также улучшение оригинала и исправление перевода, желательно с помощью автоматизированных процедур.

Несмотря на то, что наилучшие результаты возможны только при работе над всеми направлениями, в целях нашего исследования мы решили ограничиться только одним: улучшением исходного текста в соответствии с основными принципами международного англий-

ского языка. Мы использовали систему на основе правил, поскольку она более чувствительна к грамматическим улучшениям. Фактически, система на основе правил анализирует предложение, чтобы понять его. Следовательно, улучшив оригинал, тем самым мы улучшим и перевод. Но чтобы воспользоваться также преимуществом статистического машинного перевода — возможностью делать текст более складным, мы выбрали новый гибридный обработчик SYSTRAN.

Исследование

Для оценки улучшений текста оригинала мы решили использовать такой показатель, как эффективность последующего редактирования. Хотя для сравнения обученных систем целесообразно применять другие показатели, например алгоритм BLEU, нам нужен был критерий, который полностью соотносится с оценкой качества человеком, а также со скоростью и затратами.

Мы обнаружили, что эффективность последующего редактирования, т. е. среднее время, необходимое редактору для преобразования машинного перевода в полноценный «живой» текст, лучше всего соотносится с другими показателями качества, например моделью контроля качества LISA, не говоря уже о, безусловно, субъективных оценках людей. Более того, если учесть, что качество необработанного результата МП определяет скорость последующего редактирования и, соответственно, снижение расходов для заказчика, эффективность редактирования машинного перевода предоставляет ценную информацию о качестве, скорости и стоимости.

Контент для исследования был предоставлен компанией SAS Institute, крупнейшим независимым поставщиком программного обеспечения для бизнес-аналитики. Над исследованием, в частности, работали Джон Кол (John Kohl), технический редактор и инженер-лингвист компании SAS Institute, автор книги *The Global English Style Guide* (Международный английский язык. Руководство по стилю) (2008), а также Ричард Меннеглир (Richard Menneglier), руководитель проектов по локализации в парижском офисе компании Lexcelera.

Тестовый документ объемом 880 слов представлял собой часть интерактивной справки к программному обеспечению SAS Anti-Money Laundering. Этот документ выбрали потому, что он был очень хорошо написан в соответствии со стандартами, которых придерживается большинство компаний, но без учета необходимости в будущем переводе. В нем не было грамматических, орфографических или терминологических ошибок, но он нарушал ряд основных принципов международного английского языка, которые описал Джон Кол в своем руководстве по стилю. Хотя документ состоял из разделов справки, отобранный материал был достаточно содержательным, это не были инструкции по выполнению задач. Синтаксис инструкций, ориентированных на выполнение задач, был бы проще, и возможностей преобразования информации в более подходящий для системы МП вид было бы меньше.

Европейский центр локализации SAS предоставил перевод около 500 техни-

ческих терминов и названий элементов пользовательского интерфейса, встречающихся в документации к SAS Anti-Money Laundering. Джон Кол, как технический редактор, определил, что 56 из предоставленных терминов встречались в тестовом документе, а Ричард Меннеглир, руководитель проектов, использовал их как средство «мини-обучения» гибридной системы SYSTRAN. После чего технический редактор изменил текст оригинала в соответствии с правилами международного английского языка. Таким образом, мы получили две версии исходного документа: отредактированную и оригинальную. Для сравнения результатов предварительного редактирования с результатами обучения системы мы проверили отредактированный и оригинальный тексты с помощью обученной и необученной системы МП. Иными словами, мы фактически тестировали четыре сценария: необученная система МП и первоначальный исходный документ; необученная система МП и отредактированный исходный документ; обученная система МП и первоначальный исходный документ; обученная система МП и отредактированный исходный документ. Последующее редактирование каждого файла выполнялось отдельно, а затраченное на него время тщательно отслеживалось.

Результаты

Необученная система показала неудовлетворительные результаты как с первоначальным, так и с отредактированным исходным документом. Неудивительно, что самые плохие переводы получаются, когда система МП используется «как есть», без обучения. Программе трудно «понять» базовую терминологию, в результате редактору приходится тратить больше времени на исправление терминов. Кроме того, неотредактированный текст, в котором не соблюдены правила международного английского языка, машине, как и человеку, воспринимать сложнее.

Интересно, что без правильно обученной системы даже хорошо написанный текст не дает ощутимого улучшения результата. При использовании необученной системы и первоначального исходного документа производительность редактора составляла 5587 слов в день — приличные темпы, если учесть, что средняя скорость перевода специалистом — 2500 слов. Но этому результату далеко до потенциала МП. Используя необученную систему и отредактированный исходный документ, темпы удалось поднять лишь незначительно — до 6208 слов в день.

Обученная система позволила достичь пика производительности, особенно с отредактированным исходным материалом. Хотя цель данного исследования не заключалась в оценке влияния адаптации системы на результат, такая адаптация, вне всякого сомнения, приносит наиболее существенную выгоду. Как только были добавлены словари для обучения системы, значительно улучшилось качество переводов как первоначального, так и отредактированного исходного текста. Продуктивность редактора возросла до 7880 слов в день даже без оптимизации исходного текста. Эта цифра отражает существенное улучше-

ние качества на выходе, в основном благодаря наличию соответствующей терминологии в системе, что позволяет избежать трудоемкого поиска терминов — именно на это уходит больше всего времени при последующем редактировании. Тем не менее, при работе с неотредактированным исходным текстом результат по-прежнему содержал грамматические ошибки, и конечная продуктивность редактора оставляла желать лучшего.

Неудивительно, что наилучшей комбинацией для повышения эффективности редактирования стало сочетание обученной системы и оптимизированного (в данном случае предварительно отредактированного) исходного контента. При использовании такой комбинации качество перевода было очень высоким, исчезли также многие грамматические ошибки. Структура предложений в исходном тексте упростилась, что позволило системе SYSTRAN правильно обработать контент. В этом случае редакторы ограничились лишь незначительными точечными правками, улучшив предложения до уровня перевода, выполненного носителем языка. Производительность была исключительной: 9677 слов в день. Подведем итоги. Редактировать МП без обучения системы вдвое быстрее, чем переводить с нуля; с обучением — втрое быстрее; с обучением и контролем исходного текста — в четыре раза быстрее.

Значительные улучшения исходного текста

Поскольку при улучшении исходного контента повышается производительность, мы перешли ко второй цели нашего проекта — выбора из многочисленных правил международного английского языка тех, которые в наибольшей степени отражаются на качестве машинного перевода. После анализа мы определили три правила, больше всего влияющие на перевод. Исходный текст, отредактированный и нет, приведен в

Правило 1. Употребляйте глаголы в действительном залоге вместо герундия

Исходный текст (первоначальный и отредактированный)	Перевод (необработанный машинный перевод)
Первоначальный текст. Understanding the differences between owned and checked out alerts is critical to understanding SAS® Anti-Money Laundering.	La compréhension des différences entre les alertes possédées et Extraites est critique au SAS® Anti-Money Laundering de compréhension.
Отредактированный текст. In order to understand SAS® Anti-Money Laundering, you need to understand the differences between owned alerts and checked out alerts.	Afin de comprendre le SAS® Anti-Money Laundering, vous devez comprendre les différences entre les alertes détenues par un autre utilisateur et les alertes bloquées.
Обратите внимание, что при переводе исходного улучшенного текста оригинала редактору пришлось внести одно единственное изменение в машинный перевод. Afin de comprendre le fonctionnement de SAS® Anti-Money Laundering, vous devez comprendre les différences entre les alertes détenues par un autre utilisateur et les alertes bloquées.	

Правило 2. Не употребляйте пассивный залог

Исходный текст (первоначальный и отредактированный)	Перевод (необработанный машинный перевод)
Первоначальный текст. Risk-factor-only alerts can be identified by the Scenario and Triggering Values columns on an alert list window.	Des alertes de type facteur de risque uniquement peuvent être identifiées par le scénario et des colonnes Valeurs de déclenchement sur une fenêtre de listes des alertes.
Отредактированный текст. To identify a risk-factor-only alert, the Scenario column of the alert list window displays either ML_Risk or TF_Risk.	Pour identifier une alerte de type facteur de risque uniquement, la colonne Scénario de la fenêtre de listes des alertes montre ML_Risk ou TF_Risk.
Снова перевод улучшенного исходного текста потребовал минимальных правок на стадии последующего редактирования: Pour identifier une alerte de type facteur de risque uniquement, la colonne Scénario de la fenêtre de listes des alertes indique ML_Risk ou TF_Risk.	

Правило 3. Пишите короткими предложениями. Одно предложение должно содержать одну мысль

Исходный текст (первоначальный и отредактированный)	Перевод (необработанный машинный перевод)
Первоначальный текст. Alerts are displayed on alert list windows, which provide tools and information to aid users as they determine whether alerts represent suspicious activity that should be reported to authorities.	Des alertes sont montrées sur les fenêtres de listes des alertes, qui fournissent des outils et des informations aux utilisateurs d'aide pendant qu'elles déterminent si les alertes représentent l'activité suspecte qui devrait être rapportée aux autorités.
Отредактированный текст. Alerts are displayed in alert list windows. The alert list windows provide tools and information that help users determine whether alerts indicate suspicious activity that should be reported to authorities.	Des alertes sont montrées dans des fenêtres de listes des alertes. Les fenêtres de listes des alertes fournissent les outils et les informations qui aident des utilisateurs a déterminer si les alertes indiquent l'activité suspecte qui devrait être rapportée aux autorités.
Обычно это два правила в одном. Более короткие предложения в целом и ограничение количества тезисов в одном предложении улучшают результат МП. Ниже приведены правки редактора. Les alertes s'affichent dans des fenêtres de listes des alertes. Les fenêtres de listes des alertes fournissent les outils et les informations qui aident des utilisateurs a déterminer si les alertes indiquent une activité suspecte qui devrait être signalée aux autorités.	

таблице слева, а результат машинного перевода на французский язык — справа. Изменения подчеркнуты.

Мы вместе с Грегом Окстоном (Greg Oxton) из Консорциума по инновациям в области услуг считаем, что из всех возможных улучшений исходного текста для повышения его пригодности для машинного перевода наиболее действенное правило для технических писателей

— ограничение количества тезисов в предложении. У этого правила есть и дополнительное преимущество: текст, который легко понимает система МП, так же легко поймут и люди.

В заключение можно сказать, что хорошо обученная система в сочетании с исходным контентом, написанным по правилам международного английского языка, позволяет добиться самого высокого качества машинного перевода. Нашей отправной точкой в данном исследовании была гибридная система SYSTRAN, не прошедшая обучение, и даже она позволила выполнить последующее редактирование вдвое быстрее, чем перевод с нуля. Тем не менее, простая адаптация системы с внесением правильной терминологии привела к тому, что скорость последующего редактирования стала втрое выше, чем скорость перевода специалистом. Если добавить такую составляющую, как качественное написание текста, производительность редактора вырастет в четыре раза по сравнению с производительностью переводчика. Такой многообещающий результат говорит о возможной выгоде при любом контроле исходного текста, будь то управление составлением текста, например в программе acrolinx IQ; предварительное редактирование, включая стандартизацию текста, либо соблюдение лишь нескольких наиболее эффективных правил, например ограничение количества тезисов в предложении.

Кроме того, система МП на основе правил, похоже, особенно хорошо реагирует на улучшения грамматической структуры исходного текста. Использование обученной системы и качественного исходного текста позволяет повысить качество машинного перевода, а значит ускорить последующее редактирование и снизить затраты на локализацию.

Лори Тике *— соучредитель и генеральный директор компании Lexcelera, соучредитель организации «Переводчики без границ» и член редколлегии журнала MultiLingual.*