Н. В. Бабушкина

РАЗРЕШЕНИЯ ОМОНИМИИ ING-ФОРМ ПРИ МАШИННОМ ПЕРЕВОДЕ СПЕЦИАЛЬНОГО ТЕКСТА: К ВОПРОСУ О СООТНОШЕНИИ СИСТЕМА МП-РЕДАКТОР

Кафедра прикладной лингвистики филологического факультета.
Научный руководитель - Л. Н. Беляева

В настоящее время системы машинного перевода (МП) уже широко используются в крупных организациях и переводческих бюро, на порталах сети Интернет, а также в работе независимых переводчиков. Хотя такие системы не создают перевод, равный профессиональному, с их помощью можно значительно ускорить и удешевить сам процесс перевода, если подходить к вопросу оценки качества и адекватности перевода с точки зрения задач и требований заказчика или пользователя. Дело в том, что для извлечения общей информации из текста специалисту достаточно даже грубого МП, а для создания текста издательского качества такой перевод обязательно потребует редактирования[1]. При этом редактирование должно занимать меньше времени, чем ручной перевод, иначе использование системы МП становится невыгодным[2]. Следовательно, редактор должен быть подготовлен к работе с результатами МП так, чтобы редактирование из процесса быстрого исправления не превратилось в переделывание перевода.

Одной из составляющих подготовки редакторов для работы с системами МП должно быть изучение случаев, сложных для алгоритмизации, и наиболее типичных ошибок машинного перевода. Ввиду различного подхода к решению задач редак-

тирования в тексте МП следует различать неточности и ошибки: *неточности* связаны лишь со стилистической некорректностью переводного предложения и могут быть отредактированы без обращения к исходному тексту, поскольку в целом перевод понятен; *ошибки* препятствуют пониманию текста, для их редактирования зачастую необходимо заново анализировать исходное предложение или его часть.

Установление и описание сложных для МП единиц и конструкций не только укажет редактору, на что необходимо обращать внимание в тексте перевода, но и позволит определить изменения, которые нужно ввести в словарные статьи автоматических словарей (АС)[3] и алгоритмы.

Особую сложность для машинного анализа (парсинга) английского текста представляют ing-овые словоформы (ing-формы). Известно, что в английском языке существует ряд частей речи, образующихся с помощью флексии ***ing:*** отглагольное существительное, прилагательное, предлог, союз, а также глагольные формы: причастие I, инфинитив в форме Continuous и герундий. Такое многообразие ing-форм и категорий значительно затрудняет машинную обработку текстов из-за необходимости снятия их омонимии.

Сложности машинного анализа и перевода ing-форм на русский язык в настоящей статье рассмотрены на примерах из выборки специальных текстов по сейсмологии за 1994-2000 гг. общим объемом около 260 000 слов. Переводы выполнены одной из современных систем МП - системой SILOD-Windows, реализованной в виде библиотеки прикладных программ WORD+, разработанной в лаборатории машинного перевода РГПУ им. А. И. Герцена.

В результате исследования установлено, что не все ing-формы одинаково трудны для машинного анализа и перевода: наиболее сложным является герундий (в специальном тексте употребляется только в простой форме Indefinite Active). При машинном анализе герундия сложности заключаются в:

а) определении статуса герундия ввиду необходимости снятия омонимии с другими формами, в основном с причастием, так как формы герундия и причастия совпадают в парадигмах всех глаголов, имеют схожую дистрибуцию, а также чаще других ing-форм используются в тексте;

б) определении связей между герундием и его окружением: благодаря наличию именных и глагольных валентностей герундий используется в большом наборе конструкций, при этом в одном и том же синтаксическом окружении может выполнять различные функции.

Сложность перевода герундия объясняется отсутствием аналогичной формы в русском языке и многообразием герундиальных конструкций: поскольку универсального способа перевода герундия на русский язык не существует, необходимы индивидуальные модели перевода для конкретных употреблений.

Системой МП выполняется анализ на следующих уровнях: лексико-морфологическом, уровне групп, функциональных сегментов, предложения. Возникновение ошибок возможно на всех уровнях анализа исходного текста, что приводит к неверному преобразованию (трансферу) и, соответственно, к неправильному переводу. Кратко рассмотрим ошибки каждого уровня, а также возможности и варианты их исправления.

Ошибки лексико-морфологического уровня наиболее редки в переводе герундия и наиболее просты для исправления, поскольку во многих случаях достаточно изменить информацию в АС. К таким изменениям относится пополнение АС неопознанными ранее словами. Кроме того, иногда возникают ошибки в переводе членов омонимичного ряда, один из которых

иг.пг>тп,7\/ртга иятпр u p.u nnvrup T-Тяпм-

мер, словоформа ***following*** определяется системой МП только как прилагательное, хотя в приведенном ниже примере является герундием:

Rothman solved the prevalent problem of "cycle skips" in linearized methods of residual

statics computation of noisy data by following the Monte Carlo optimization technique of Metropolis et al... - Rothman решил распространенную проблему «рикошеты цикла» в линеаризованный методы остаточного вычисления statics шумных данных путем следующего техники оптимизации в Монте Карло Метрополис идр...

Для таких случаев какие-либо изменения в алгоритме для улучшения перевода сделать трудно, поскольку невозможно выделить однозначные формальные показатели прилагательного и герундия: у них схожая дистрибуция (предлог в препозиции, существительное в постпозиции и т. п.). Так как герундий *following* используется в тексте значительно реже омонимичного прилагательного, то введение дополнительного «герундиального» перевода будет лишь загромождать текст. Зная о возможной ошибке, редактор может исправить перевод герундия *following,* не обращаясь к исходному предложению, поскольку его значение понятно.

Самое большое количество ошибок парсинга приходится на уровень групп: в английском языке нет развернутой падежной системы и категории грамматического рода, возможна конверсионная омонимия, сильна беспредложная связь и т. п. - иными словами, набор формальных показателей, на которые система МП опирается при анализе групп, сравнительно невелик. Основные сложности возникают при определении границ групп, семантических связей между элементами и, при наличии омонимов, - статуса омонимов. Анализ выборки показал, что ошибки в определении статуса герундия возникают при переводе простых именных групп, неточности чаще встречаются при переводе групп с предложным управлением.

Под группами с предложным управлением понимаются группы, в которых предлог относится к ing-форме и стоит непосредственно перед ней. В исследованной выборке наиболее часто герундий употребляется с предлогами: *by, for, in* и *of.* В таких группах система МП практически всегда верно определяет статус герундия: поскольку основные трудности вызывает снятие омонимии типа герундий/причастие, при переводе предложных конструкций предлог работает как индикатор «не-причастия». Связи внутри группы также устанавливаются системой МП верно, так как зависимость между элементами, как правило, линейная: следующий элемент зависит непосредственно от предыдущего, что является наиболее простой схемой для машинного анализа.

В сочетании с некоторыми предлогами (в выборке: *behind, into, upon* и *on)* перевод герундия может быть неточен и требует редактирования. Например:

The motivation behind resorting to the inversion scheme to be dealt with in this paper comes from the absence of even a one-dimensional local velocity model... - Мотивация за обращением к схеме инверсии, которую нужно рассмотреть в этом докладе исходит из отсутствия даже одномерная местная скоростная модель...

Слово *мотивация* в русском языке не принимает зависимых слов с предлогом за, а управляет существительным в родительном падеже без предлога. Поскольку перевод понятен, для редактирования не требуется обращения к исходному предложению.

Под простыми именными герундиальными группами (ИГГ) понимаются именные группы из двух и более простых элементов, в которых ing-форма не управляется с помощью предлога, непосредственно стоящего перед ней, и связи между членами группы не оформлены предлогами. Такие группы различаются по количеству элементов и по позиции герундия в группе. В текстах наиболее распространены двух- и трехэлементные группы, группы с пятью и более составляющими встречаются очень редко. При этом с увеличением длины группы растет процент ошибочных переводов, так как структура зависимостей между элементами становится более сложной.

В большинстве случаев ядром простой ИГГ является ее последний элемент, это

чаще всего не ing-форма (за исключением двухкомпонентных групп), герундий употребляется непосредственно слева от ядра, т. е. в предпоследней позиции в группе, а зависимости в группе направлены справа налево. Ошибки в переводе чаще всего заключаются в том, что герундий определяется системой МП как причастие, элемент, стоящий непосредственно перед герундием, - как имя, управляющее причастием, а, соответственно, связи в группе-как направленные слева направо. В русском языке такой перевод реализуется причастным оборотом:

*The aim of our depth **imaging** methodology is to determine the geometry and the location of fault planes... - Цель нашей глубины, отображающей методологию, состоит в том, чтобы определить геометрию и расположение плоскостей сброса...*

Система МП правильно производит анализ лишь некоторых двухкомпонентных групп, а также тех трехкомпонентных групп, где перед герундием стоит словоформа, которая не может принимать причастный оборот, как *painstaking* в следующем примере:

*Setting out each receiver, however, required a painstaking **leveling** process... - Выставляющий каждый приемник, однако, необходимый тщательный процесс нивелирования...*

Улучшить перевод простых ИГГ посредством изменения алгоритма затруднительно. Ошибки можно исправлять только редактированием, причем потребуется повторный анализ исходного предложения, поскольку в большинстве случаев по переводу невозможно определить границы проблемной группы и связи между ее элементами. Устойчивые сочетания, образующие группу или ее часть, следует вводить в АС как иконические обороты, чтобы избежать ошибок в их переводе.

[illegible]
[illegible]
ментов определяется герундий в следующих функциях: подлежащего, части сказуемого (глагольного и именного) и беспредложного дополнения. В составе указанных сегментов герундий в основном употребляется следующим образом: герундиальное подлежащее - в абсолютном начале предложения, герундий в составе сказуемого - непосредственно после вспомогательного глагола, беспредложное дополнение - непосредственно после сказуемого в активном залоге и в постпозиции с существительным/именной группой с артиклем или предлогом. Такие употребления хорошо переводятся системой МП. Следующий пример иллюстрирует перевод беспредложного дополнения.

*The physics of rock stability may allow **increasing** the minimum magnitude to 6.0. - Физика устойчивости породы может позволять увеличение минимальной величины 6.0.*

Если употребление герундия в составе функциональных сегментов отлично от вышеописанных, то возможны ошибки в переводе. Например, герундиальное подлежащее после вводного оборота иногда переводится неверно, так как система МП может не распознать этот оборот как иконический:

*In the present case, **depending** only on the average error criterion seems to be sufficient. - В присутствующем случае [корпусе], зависящем только в среднем, критерий ошибки, как кажется, будет достаточным.*

В данном случае оборот *in the present case* следует ввести в АС.

В целом редкие ошибки в переводе функциональных сегментов могут быть исправлены редактированием, в некоторых случаях даже без повторного анализа исходного текста.

Сложности анализа и перевода герундия на уровне предложения в обработанном массиве вызывали только однородные члены. В большинстве случаев система МП не распознает связи между ними, независимо от того, образованы они герундием или другими словоформами: система МП анализирует не связь между однородными членами, а **ГруП-**
лит., [illegible]
дельно, что приводит к ошибкам в переводе простых именных групп. Так, например, статус второго герундиального подлежащего в следующем случае определен неверно:

***Picking** fault locations and **interpreting the** core of a fault-propagation fold are problematic*

because of large lateral velocity changes. - Выбор обнаружений неисправности и при интерпретации активную зону распространения складки разлома являются проблематичными из-за больших боковых скоростных изменений.

Редактирование перевода беспредложных однородных членов всегда требует повторного анализа исходного предложения, поскольку в переводе часто трудно выделить структуру самого предложения, однородные члены и композицию группы каждого из них.

По результатам проведенного исследования можно сделать вывод, что наибольшее число ошибок машинного анализа приходится на уровень групп и уровень предложения. В переводе герундия в составе простых именных групп и однородных членов ошибки значительно преобладают над неточностями и верным переводом, следовательно, для редактирования практически всегда необходимо заново производить анализ исходного предложения. Редактору, знакомому с описанными в настоящей статье типичными композициями сложных для перевода групп, распределением зависимостей в группе и наиболее распространенными ошибками анализа и перевода, будет легче выделить в тексте перевода и оригинала и отредактировать проблематичную группу.

По результатам исследования переводов текстовой выборки можно также порекомендовать авторам текстов для последующего машинного перевода по возможности не использовать простые именные группы с числом элементов больше трех и развернутые предложения с несколькими беспредложными герундиальными однородными членами.

ПРИМЕЧАНИЯ

[1] См., например, FEMTI <http://www.issco.unige.ch:8080/cocoon/femti/st-home.html>; *Hutchins W. J.* Machine Translation and Human Translation: in Competition or in Complementation? // Machine Translation: Theory & Practice. —-New Delhi: Bahri Publications, 2001. — P. 5-20.

Popescu-Belis A. An Experiment in Comparative Evaluation: humans vs. computers [Electronic source] // Proceedings of MT Summit IX.—New Orleans, Louisiana, USA, 2003 <http://www.amtaweb.org/summit/MTSummit/FinalPapers/60-Popescu-final.pdf>

[2] *Hutchins W. J., Somers H. L.* An Introduction to Machine Translation. — London; New York: Academic Press, 1992. — P. 152-153.

[3] *Correa N.* A Fine-grained: Evaluation Framework for Machine Translation System Development. [Electronic source] // Proceedings of MT Summit IX. — New Orleans, Louisiana, USA, 2003 <http://www.amtaweb.org/summit/MTSummit/FinalPapers/56-Correa-final.pdf>