СТАТИСТИЧЕСКИЕ И ГИБРИДНЫЕ МЕТОДЫ ПЕРЕВОДА В ТЕХНОЛОГИЯХ КОМПАНИИ PROMT

АЛЕКСАНДР МОЛЧАНОВ
Alexander.Molchanov@promt.ru

Машинный перевод, существующий уже несколько десятилетий, в последние годы переживает бурный рост, главным образом за счет использования статистических технологий. Чем различаются системы машинного перевода и какие процессы характерны для них в настоящее время, показано в данной статье.

ОСНОВНЫЕ ТИПЫ СИСТЕМ МАШИННОГО ПЕРЕВОДА

История машинного перевода начинается с так называемого «Джорджтаунского эксперимента». В январе 1954 г. в Нью-Йорке состоялась первая публичная демонстрация системы машинного перевода с русского языка на английский, разработанной компанией IBM совместно с Джорджтаунским университетом. Система по современным меркам была примитивной и включала в себя словарь объемом 250 слов и грамматику из шести правил. Эксперимент получил широкий резонанс, и исследования в области разработки систем машинного перевода начались по всему миру, в том числе и в СССР.

В 1966 г. созданная правительством США комиссия ALPAC (Automatic Language Processing Advisory Committee) опубликовала печально известный доклад, согласно которому разработка систем машинного перевода была признана нерентабельной. Это фактически привело к повсеместному прекращению работ над системами машинного перевода. Однако благодаря постоянному прогрессу вычислительной техники исследования в этой области вновь возобновились в 70-е годы, а в конце 80-х начинается разработка первых статистических систем.

Уже в 1980-е сложился рынок коммерческих разработок систем машинного перевода. По данным агентства WinterGreen Research, в 2012 г. мировой рынок машинного перевода составлял $1,6 млрд, а к 2019 г., как ожидается, достигнет $6,9 млрд. В настоящее время существует множество компаний, которые занимаются коммерческой разработкой систем машинного перевода: SYSTRAN, PROMT, Linguatec, Asia Online, Safaba и др.

Целью использования машинного перевода может быть как получение перевода высокого качества, так и простая передача смысла исходного текста (так называемый «джистинг»). Машинный перевод применяется для перевода следующих типов текста:

- пользовательский контент (отзывы, комментарии и т. д.);
- документация (техническая, эксплуатационная, юридическая и т. д.);
- новостной контент;
- каталоги интернет-магазинов;
- личная и деловая переписка.

К основным сферам применения машинного перевода относятся:

- локализация (ускорение и удешевление перевода больших объемов текста, например документации к ПО);
- оптимизация работы переводчиков и переводческих бюро (результат машинного перевода редактируется переводчиками);
- Интернет (электронная торговля, новостные и образовательные сайты).

В настоящее время существует два основных типа систем машинного перевода: основанные на правилах (rule-based machine translation, RBMT) и статистические.

СИСТЕМЫ, ОСНОВАННЫЕ НА ПРАВИЛАХ

В системах, основанных на правилах, можно выделить два основных подтипа: трансферные и системы-интерлингвы.

Трансферные системы машинного перевода распространены более широко, чем системы-интерлингвы. Они работают по следующим принципам: проводится морфологический, лексический и семантико-синтаксический анализ предложения на языке оригинала, создается синтактико-семантическое дерево разбора входного предложения, затем производится так называемый «трансфер», т. е. преобразование структуры входного предложения в соответствии с формальными требованиями языка перевода. На заключительном этапе синтеза формируется конечное предложение на языке перевода. Основанная на правилах система перевода PROMT является классическим примером трансферных систем.

В основе **систем-интерлингв** лежит теория о том, что любое предложение любого языка можно преобразовать в его смысловое представление на так называемом универсальном метаязыке. А из полученного смыслового представления можно синтезировать предложение на языке перевода. Иными словами, с помощью определенного набора правил и словаря с семантическими характеристиками можно преобразовывать текст в смысл и наоборот. Интерлингвы требуют очень долгой разработки и создания огромных баз знаний о языке.

Системы, основанные на правилах, обладают рядом общих характеристик. С точки зрения устройства, они включают в себя словари и формальные грамматики, т. е. наборы правил морфологического, семантического и синтаксического анализа языка. С точки зрения разработки и эксплуатации, такие системы обладают рядом преимуществ и недостатков.

Достоинства: высокое качество, стабильность и предсказуемость машинного перевода.

Недостатки: высокая стоимость разработки и поддержки лингвистических алгоритмов и словарей, а также большое количество времени, необходимое для лексической настройки системы для отдельного клиента или новой предметной области. Кроме того, при высокой точности основанный на правилах перевод обладает определенным «машинным» акцентом, т. е. часто выглядит неестественно.

Современные RBMT-системы обычно включают в себя общетематические словари (объемом от нескольких десятков до нескольких сотен тысяч статей) и специализированные словари по отдельным тематикам (объемом до нескольких десятков тысяч статей).

ТАБЛИЦА 1. ОБЪЕМЫ ОБЩЕТЕМАТИЧЕСКИХ СЛОВАРЕЙ RBMT-СИСТЕМЫ PROMT ДЛЯ ОСНОВНЫХ НАПРАВЛЕНИЙ ПЕРЕВОДА

Направление	Объем словаря (тыс. статей)
Англо-русское	220
Англо-французское	67
Англо-немецкое	81
Англо-итальянское	61
Англо-португальское	70
Англо-испанское	82

В таблице 1 приведены статистические данные по объему общетематических словарей RBMT-системы PROMT.

Производительность RBMT-систем машинного перевода зависит от различных параметров (среди которых количество и сложность грамматических правил, объем и количество используемых словарей) и обычно варьируется от нескольких слов до нескольких сотен слов в секунду. Например, производительность RBMT-системы PROMT для англо-русского направления составляет примерно 150–200 слов в секунду при переводе в один поток на компьютере с процессором Intel® Core™ i7-2600K CPU с частотой 3,40 ГГц.

СТАТИСТИЧЕСКИЕ СИСТЕМЫ

В основе любой системы статистического машинного перевода лежит использование массивов текстов, представленных одновременно на языке оригинала и языке перевода. Такие массивы данных называются параллельными корпусами текстов. Сначала статистическая система проходит этап обучения, на котором извлекаются статистические данные о переводе отдельных слов и фраз с исходного языка на язык перевода. В процессе перевода такая система вычисляет наиболее вероятный перевод исходного предложения на основе данных, полученных при обучении. Помимо параллельного корпуса текстов, статистические системы используют корпусы текстов на языке перевода. На основе такого корпуса строится статистическая модель языка перевода, которая используется при оценке того, насколько вариант перевода предложения адекватен и «гладок» с точки зрения норм и правил языка перевода.

Достоинства: быстрая настройка (по сравнению с системами, основанными на правилах), самообучаемость (участие эксперта при настройке системы можно свести к минимуму), а также высокая «гладкость» перевода (перевод очень похож на человеческий и в нем практически отсутствуют шероховатости).

Недостатки: необходимость наличия качественных параллельных корпусов большого объема для настройки системы. Кроме того, статистический перевод часто содержит большое количество грамматических ошибок (особенно когда речь идет о языках с богатой морфологией, таких как, например, русский или немецкий) и в целом отличается нестабильностью и непредсказуемостью (к примеру, одна и та же конструкция может переводиться совершенно по-разному в разных контекстах, в переводе могут пропадать слова и т. д.).

Производительность современных статистических систем может широко варьироваться и зависит, в первую очередь, от трех факторов:

- объем модели перевода;
- объем языковой модели;
- объем выделяемой оперативной памяти.

Производительность статистической системы PROMT для модели перевода объемом около 100 млн. словоупотреблений составляет 15–20 слов в секунду при переводе в один поток на компьютере с процессором Intel® Core™ i7-2600K CPU с частотой 3,40 ГГц и объемом оперативной памяти 16 Гбайт.

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ В СИСТЕМЕ PROMT

Компания PROMT занимается разработкой статистических и гибридных систем машинного перевода с 2008 г. Основной мотивацией для такой работы послужило, с одной стороны, стремление преодолеть недостатки RBMT-системы за счет создания гибридной системы перевода, а с другой стороны — создание статистических систем для тех языковых пар, которых нет в базовой системе PROMT. Так, компания занимается разработкой статистического перевода для казахского, финского, китайского, японского и скандинавских языков.

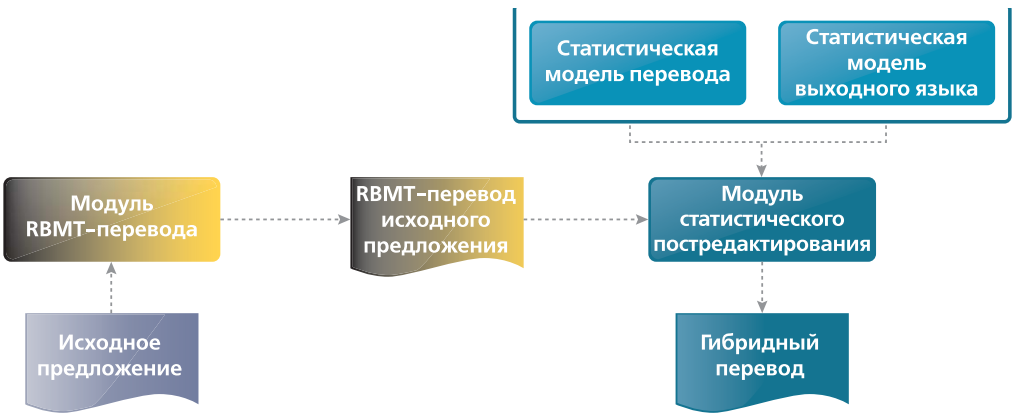
ПРИНЦИПЫ РАБОТЫ ГИБРИДНОЙ СИСТЕМЫ PROMT

Гибридное решение PROMT доступно для всех языков базовой системы, которые включают в себя русский и основные европейские языки. В основе гибридной системы PROMT лежит идея о том, что с помощью параллельного корпуса текстов и специального статистического модуля можно, во-первых, быстро и качественно настроить перевод для определенной предметной области, а во-вторых, исправить недостатки, ошибки и шероховатости перевода, основанного на правилах. Такой специальный модуль называется модулем статистического постредактирования.

Гибридная система, так же как и статистическая, проходит процесс обучения на параллельных данных.

РИС. 1. Схема обучения гибридной системы перевода

РИС. 2. ►
Процесс перевода предложения гибридной системой



Обучение можно разделить на три стадии:

1. осуществляется перевод исходной части параллельного корпуса на языке оригинала базовым RBMT-модулем перевода;
2. настраивается статистическая модель перевода с «машинного» языка на человеческий;
3. настраивается статистическая модель на основе корпуса языка перевода.

Схема обучения гибридной системы представлена на рис. 1.

Гибридная система PROMT содержит два основных компонента: базовый RBMT-модуль перевода и модуль статистического постредактирования, который использует данные, полученные на этапе обучения (статистическая модель перевода, статистическая модель выходного языка). В процессе перевода сначала исходное предложение переводится базовым модулем, затем полученный перевод обрабатывается статистическим компонентом, т. е. фактически на этом этапе осуществляется перевод с «машинного» языка на человеческий по правилам статистического машинного перевода. Схема процесса перевода гибридной системой представлена на рис. 2.

ОЦЕНКА КАЧЕСТВА ПЕРЕВОДА ГИБРИДНОЙ СИСТЕМЫ PROMT

Многие исследователи говорят о способности гибридных систем опережать по качеству перевода как RBMT-системы, так и статистические. К примеру, разработчики компании SYSTRAN в статье Statistical Post-Editing on SYSTRAN's Rule-Based Translation System отмечают, что их гибридная система перевода с модулем постредактирования превосходит базовую RBMT-систему.

Специалисты компании PROMT провели серию экспериментов по сравнению RBMT-, статистической и гибридной системами для англо-русского направления перевода. Эксперименты проводились на текстах компании PayPal, которая является клиентом компании PROMT. Тексты представляют собой английское руководство по использованию сервисов PayPal и его локализованную русскую версию.

Объем корпуса для обучения гибридной и статистической систем составил примерно 1 млн словоупотреблений. Тестирование систем проводилось на выборочной совокупности из ста случайным образом отобранных из обучающего корпуса предложений. При тестировании использовалась экспертная лингвистическая и автоматическая оценки на основе метрики BLEU (Bilingual Evaluation Understudy). Гибридная система сравнивалась с RBMT-системой, а также со статистической системой PROMT, настроенной на текстах PayPal. Кроме того, в сравнение был включен статистический перевод с онлайн-сервиса Google Translate.

Метрика BLEU была разработана сотрудниками компании IBM и является одной из самых простых и популярных метрик оценки машинного перевода. Алгоритм BLEU оценивает качество перевода по шкале от 0 до 100 на основании сравнения машинного перевода с человеческим и поиска общих слов и фраз. Основная идея разработчиков метрики состоит в том, что чем лучше машинный перевод, тем больше он должен быть похож на человеческий. Результаты автоматической оценки представлены в таблице 2.

Также была проведена экспертная оценка. Перевод гибридной системы попарно сравнивался с переводами других систем в терминах «лучше» (один из переводов явно превосходит другой по качеству) и «эквивалентно» (два перевода принципиально не отличаются друг от друга по качеству). При оценке учитывались грамматическая и лексическая правильность, адекватность (правильная передача смысла исходного текста) и гладкость перевода.

ТАБЛИЦА 2. РЕЗУЛЬТАТЫ АВТОМАТИЧЕСКОЙ ОЦЕНКИ МАШИННОГО ПЕРЕВОДА ВЫБОРОЧНОЙ СОВОКУПНОСТИ ИЗ КОРПУСА PAYPAL ДЛЯ РАЗЛИЧНЫХ СИСТЕМ С ПОМОЩЬЮ МЕТРИКИ BLEU

Система перевода	Значение BLEU
Гибридная система PROMT	29,2
RBMT-система PROMT	16,7
Статистическая система PROMT	27,3
Google Translate	15,2

Результаты экспертной оценки представлены в виде графика на рис. 3.

Результаты экспериментов показывают, что гибридная система превосходит RBMT- и статистическую систему согласно как автоматической, так и экспертной оценке.

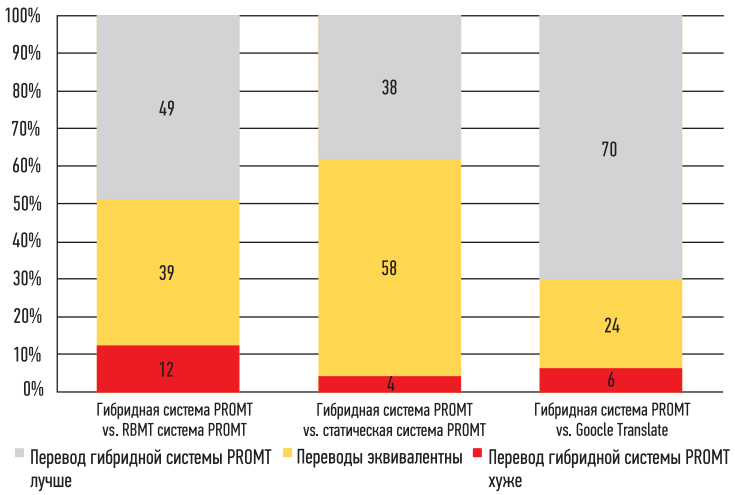


РИС. 3. ◀ Результаты экспертной оценки машинного перевода выборочной совокупности из корпуса PayPal для различных систем

ПРОБЛЕМЫ ПРИ ИСПОЛЬЗОВАНИИ СТАТИСТИЧЕСКИХ ТЕХНОЛОГИЙ В СИСТЕМЕ МАШИННОГО ПЕРЕВОДА

Использование статистических технологий сопряжено с рядом сложностей. Они касаются как внутренних (ухудшение качества и стабильности перевода), так и внешних факторов (поиск данных для обучения систем).

СТАБИЛЬНОСТЬ И КАЧЕСТВО ПЕРЕВОДА

В ходе разработки гибридной системы машинного перевода специалисты столкнулись с тем, что статистический компонент в некоторых случаях может привносить в базовый перевод не только улучшения, но и ухудшения. Прежде всего, это касается перевода именованных сущностей, т. е. специальных типовых языковых конструкций (даты, адреса, имена, названия организаций, числовые последовательности и т. п.). Перевод таких конструкций чрезвычайно важен для клиентов компании. К примеру, сумма и адрес юридического лица, прописанные в контракте, должны остаться такими же и в переводе этого контракта. Для решения этой проблемы статистический компонент гибридной системы PROMT использует метаинформацию, которую он получает из базового компонента. Все специальные конструкции на этапе перевода базовым модулем резервируются, т. е. помечаются специальными тегами. Статистический компонент использует данную метаинформацию и оставляет такие конструкции без изменений.

ДАННЫЕ ДЛЯ ОБУЧЕНИЯ СТАТИСТИЧЕСКИХ И ГИБРИДНЫХ СИСТЕМ ПЕРЕВОДА

Для настройки гибридной или статистической системы машинного перевода необходимы параллельные корпусы текстов достаточно большого объема (около одного миллиона словоупотреблений для гибридной системы и на порядок больше для статистической). И здесь разработчики сталкиваются с проблемой: где брать эти данные? Для настройки системы перевода для клиента используются параллельные тексты, накопленные им в ходе экспертного перевода клиентских данных переводческими агентствами.

Какие же данные использовать для настройки универсальных систем перевода? Существуют параллельные корпусы в открытом доступе. В качестве примера можно привести корпус протоколов заседаний Европарламента (доступен на двадцати языках, объем корпуса от десяти до пятидесяти миллионов словоупотреблений для каждого языка), корпус протоколов заседаний ООН (семь официальных языков ООН, объем корпуса в среднем от ста до двухсот миллионов словоупотреблений для каждого языка), корпус субтитров к различным кинофильмам (тридцать языков). Однако такие корпусы относятся к очень специфичной предметной области и подходят фактически только для перевода подобных текстов. Другими словами, среднестатистический посетитель онлайн-сервиса перевода вряд ли станет переводить протоколы заседаний какого-либо международного правового или законодательного органа.

Другой источник параллельных данных — открытые многоязычные интернет-ресурсы, например новостные порталы. В этом случае мы имеем дело с условно-параллельными данными (к примеру, новость на английском языке может иметь вольный перевод на русский или вообще не иметь его). Такие данные необходимо выравнивать, т. е. выделять среди большого объема данных действительно параллельные предложения на разных языках. Компания PROMT успешно использует технологии автоматической обработки и выравнивания условно-параллельных данных из интернет-источников для создания параллельных корпусов для различных предметных областей.

ПЕРСПЕКТИВЫ

Несмотря на существенное улучшение качества при переходе от RBMT-системы перевода к гибридной, ряд важных проблем остаются нерешенными. Так, перевод с использованием статистического компонента может содержать грамматические ошибки, которых нет в RBMT-переводе. Лингвистический отдел компании PROMT занимается поиском решения этой проблемы. Одним из возможных подходов является использование дополнительной информации из RBMT-компонента и постобработка гибридного перевода парсерами PROMT для выявления и исправления ошибок.

Также в скором времени планируется внедрить использование статистических технологий на онлайн-сервисе перевода компании www.translate.ru. ⬣

Данная статья основана на докладе автора, прозвучавшем на конференции AINL в Санкт-Петербурге 18 мая 2013 г.