

ДВОЙНОЙ ПОДХОД

Бесплатный переводчик Google и коммерческий PROMT. Две программы — две технологии перевода. Какая из них лучше, мы разберемся в этой статье.

В этой статье

Статистические и лингвистические технологии перевода

Сравнение Google Translate и PROMT 8.0

Программа-переводчик давно уже стала необходимой составляющей «джентльменского набора» для многих пользователей. Интерес к машинному переводу постоянно растет, практически прямо пропорционально росту объема информации на иностранных языках. Интернет-сайты, электронная почта, различные описания и инструкции на иностранных языках — все это хотелось бы переводить быстро и качественно. Именно для решения этих задач и существуют программы-переводчики. Тут уместно будет заметить, что часто возникает путаница, и переводчиком называют любую программу, которая осуществляет перевод слов. Уточним, что программа-переводчик позволяет переводить тексты целиком в отличие от словарей, которые предлагают только перевод отдельных слов.

О достоинствах отдельных программ было написано немало. В этой статье мы сравним машинные переводчики с точки зрения технологии. Сегодня в мире существуют две технологии машинного перевода: традиционная, на основе правил (rule-based machine translation), и статистическая (statistical-based machine translation). Рассмотрим их плюсы и минусы, чтобы определить, какая предоставляет более качественный перевод.

GOOGLE TRANSLATE

Статистический перевод

Поисковая система Google хорошо известна интернет-пользователям не только благодаря мощному поисковому механизму, но и различным дополнительным сервисам (наиболее популярные сервисы Google перечислены во врезке). Одним из них является онлайн-переводчик. Этот сервис реализован на основе технологии статистического перевода. Надо заме-

тить, что эта разработка используется только для тех направлений перевода, которые на сайте помечены как бета-версии. Слово «статистика» в названии технологии сразу наводит на размышления о том, что в системе применены математические методы для получения перевода. Точнее сказать, весь принцип работы основан на статистическом вычислении вероятности совпадений фраз из исходного текста с фразами, которые хранятся в базе системы перевода.

Для функционирования программе нужна база текстов на исходном языке и их переводов (база параллельных текстов), которую программа анализирует на основе определенных алгоритмов. Google Translate вычисляет наиболее вероятную последовательность слов выходного языка.

Плюсом статистической технологии является отсутствие необходимости разработки лингвистических алгоритмов в принципе. Как постулируют разработчики, при наличии свода статистических алгоритмов и базы парал-

Сервисы Google и онлайн-ый PROMT

Все сервисы Google можно разделить на три группы:

Поиск информации

- **Поиск в Интернете** (www.google.ru) — ключевой сервис Google. Позволяет находить различную текстовую информацию. Можно загрузить плагин Google Toolbar, и тогда окно ввода иско-мой информации всегда будет под рукой.
- **Картинки** (<http://images.google.ru>) — поиск картинок в Интернете.
- **Карты** (<http://maps.google.ru>) — очень хорошо сделанные карты горо-дов. Поиск адреса и маршрута для пере-движения.
- **Новости** (<http://news.google.ru>) — поиск по ключевым словам в новостях.
- **Поиск по блогам** (www.google.ru/blogsearch) — что пишут в блогах.
- **Desktop** (<http://desktop.google.com>) — искать можно не только в Интернете. Специальная программа по-зволяет находить файлы на собственном компьютере, причем значительно эффек-тивнее, чем «родная» функция Windows.

Общение

- **Gmail** (<http://mail.google.com>) — бесплатный почтовый сервис. Его уже мож-но считать культовым, поскольку Google первым предложила практически неограни-ченный размер почтового ящика (сейчас 2,7 Гбайт) и полную защиту от спама.
- **Blogger** (www.blogger.com) — сер-вис для создания собственного блога.
- **Talk** (www.google.com) — сервис для голосового общения через Сеть.
- **Группы** (<http://groups.google.ru>) — форумы и списки рассылок.

Полезные программы

- **Документы и таблицы** (<http://docs.google.com>) — зачем покупать дорогостоя-щий офисный пакет, если можно пользо-ваться бесплатным сервисом в Интернете?
- **Picasa** (<http://picasaweb.google.ru>) — программа для просмотра, редак-тирования и публикации изображений в Интернете.
- **Календарь** (<http://www.google.com/calendar>) — сервис для планиро-вания дел и встреч.

- **Переводчик** (<http://translate.google.com>) — 25 направлений пере-вода, из них два предназначены для русского языка.

PROMT

- Переводчик PROMT также доступен в виде бесплатного сервиса на сайте **Translate.ru**: 24 направления перевода, из них девять — с русским языком. Сайт предлагает следующие сервисы.
- **Перевод текста** — можно не только перевести, но и проверить орфографию (онлайн-ый переводчик работает на пре-дыдущей версии движка PROMT 7.0).
 - **Перевод веб-страниц** — переводит-ся не только текст, но и гиперссылки.
 - **Перевод e-mail** — письмо можно пе-ревести и сразу же послать.
 - **Электронный словарь** — большой объем слов по разным предметным областям.
 - **WAP-переводчик** — для мобильных пользователей.
 - **Переводчик для КПК** — специально адаптирован под разрешение КПК.

лельных текстов можно сделать систе-му перевода с любого языка.

Минусом статистических систем явля-ется отсутствие учета грамматических правил входного и выходного языков. В результате вместо связного перевода переводчик могут получиться совер-шенно не согласованные предложения с разрушенной структурой.

Вторая проблема заключается в том, что для работы такая система должна иметь доступ к очень большому количе-ству параллельных текстов. Очевидно, что практически единственным и совер-шенно неисчерпаемым источником такой базы может служить Интернет. Именно этим можно объяснить, что в сравнительных тестах по переводу с арабского и китайского на английский, проведенных Национальным институ-том науки и технологий США в 2005 году, именно онлайн-ый переводчик Google Translate одержал победу.

PROMT 8.0

Перевод по правилам

Если метод перевода, основанный на статистике, называется статистиче-ским переводом, тогда перевод с ис-пользованием лингвистических пра-

вил назовем переводом по правилам. Эта технология существует с начала 50-х годов прошлого века и по праву может называться традиционной. На заре компьютерной эры отсутствовали как базы параллельных текстов, так и машинные ресурсы для их обработки. Поэтому на протяжении нескольких десятков лет развивались только тех-нологии перевода на основе лингви-стических алгоритмов. По такому же принципу работают системы PROMT, Systran, Linguatex и многих других раз-работчиков. Поскольку нас интересует перевод с иностранных языков на рус-

ский, возьмем для сравнения систему PROMT 8.0 российского разработчика, компании «ПРОМТ».

Во главе угла этой технологии стоят лингвистические алгоритмы. Сначала система анализирует исходный текст: проводит морфологический анализ слов в предложении (определяет род, число, лицо и другие морфологические характеристики), затем синтаксический (определяет члены предложения: под-лежащее, сказуемое, дополнения, об-стоятельства) и на заключительном этапе выполняет синтез предложений в переведенном тексте. →

БЕСПЛАТНО И ВСЕГДА ПОД РУКОЙ Google Translate переводит как фрагменты текста, так и целые сайты. Направления перевода, в которых используется статистический метод, помечены как «БЕТА»

Таблица 1

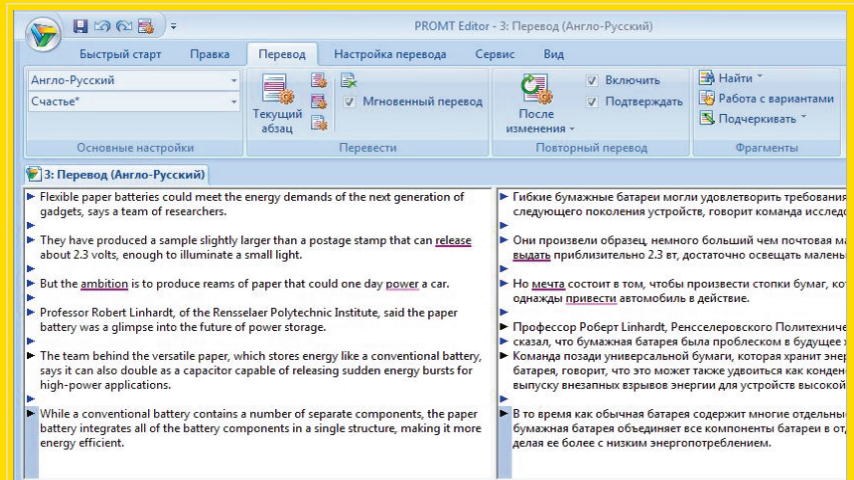
Исходный текст BBC	Перевод Google	Перевод PROMT 8.0	Перевод PROMT 8.0 после настройки
Flexible paper batteries could meet the energy demands of the next generation of gadgets, says a team of researchers.	Гибкие батареи документ может удовлетворить энергетические потребности следующего поколения гаджеты, утверждает группа исследователей.	Гибкие бумажные батареи могли удовлетворить требованиям энергии следующего поколения устройств, говорит команда исследователей.	Гибкие бумажные батареи могли удовлетворить требованиям энергии следующего поколения устройств, говорит команда исследователей.
They have produced a sample slightly larger than a postage stamp that can release about 2.3 volts, enough to illuminate a small light.	Они дали образец немного больше, чем почтовая марка, что может высвободить около 2,3 вольт, что достаточно для освещения небольшой огонь.	Они произвели образец, немного больший чем почтовая марка, которая может выпустить приблизительно 2.3 вт, достаточно освещать маленький свет.	Они произвели образец, немного больший чем почтовая марка, которая может выдать приблизительно 2.3 вт, достаточно освещать маленький свет.
But the ambition is to produce reams of paper that could one day power a car.	Но мечта — производить пачек бумаги, которые могут в один прекрасный день власть автомобиля.	Но амбиция состоит в том, чтобы произвести стопки бумаг, которые могли однажды привести автомобиль в действие.	Но мечта состоит в том, чтобы произвести стопки бумаг, которые могли однажды привести автомобиль в действие.
Professor Robert Linhardt, of the Rensselaer Polytechnic Institute, said the paper battery was a glimpse into the future of power storage.	Профессор Роберт Линхардт, из Ренсселеровского политехнического института, сказал, что документ аккумулятор взглянуть на будущее мощности.	Профессор Роберт Linhardt, Политехнического института Rensselaer, сказал, что бумажная батарея была представлением в будущее хранения мощи.	Профессор Роберт Linhardt, Ренсселеровского Политехнического института, сказал, что бумажная батарея была прорывом в будущее хранения энергии.
The team behind the versatile paper, which stores energy like a conventional battery, says it can also double as a capacitor capable of releasing sudden energy bursts for high-power applications.	Группа за универсальный документ, который хранит энергии, как обычные батареи, что она может также выступать в качестве конденсатора способное выпускать внезапной энергии очередей на высоком применения.	Команда позади универсальной бумаги, которая хранит энергию как обычная батарея, говорит, что это может также удвоиться как конденсатор, способный к выпуску внезапных взрывов энергии для мощных применений.	Команда позади универсальной бумаги, которая хранит энергию как обычная батарея, говорит, что это может также удвоиться как конденсатор, способный к выпуску внезапных взрывов энергии для устройств высокой мощности.

Результат перевода новости BBC с помощью систем Google и PROMT

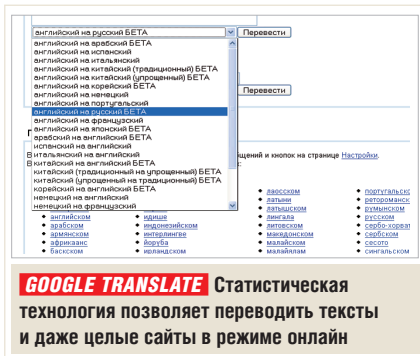
При чтении перевода Google смысл документа уловить можно, но взгляд постоянно цепляется за такие несуразности, как «которые могут в один прекрасный день власть автомобиля». Общее впечатление от перевода таково: текст склеен из кусочков фраз, где-то даже удачно, но в большинстве предложений лексическая связка между словами отсутствует. Перевод PROMT также не лишен «шершавостей», таких как «достаточно освещать маленький свет», но в целом читается легко и по качеству вполне приемлем для первого знакомства с текстом. В таблице также приведен результат перевода PROMT после проведения настройки системы (были подобраны более корректные варианты перевода для слов, выделенных красным). В системе есть целый ряд инструментов для повышения качества перевода специализированных текстов, подобных переведенному. Это особенно полезно при

регулярном переводе текстов определенной тематики. Для второго теста (см. таблицу 2) возьмем текст стандартного юридического

документа. При испытании мы использовали текст контракта, который подписывается сотрудниками при найме на работу в Университет Пенсильвании.



СОВРЕМЕННОСТЬ И ГИБКОСТЬ НАСТРОЕК PROMT 8.0 обладает интерфейсом в стиле Vista. При этом можно улучшить правильность перевода, подключив дополнительные словари



Плюсом такой системы является ее самодостаточность, ведь для работы переводчика не требуется доступ к базам параллельных текстов. Кроме того, система позволяет выполнять настройку, что серьезно повышает качество перевода специализированных текстов.

Минусом этой технологии можно назвать большую ресурсоемкость работы по улучшению качества перевода — ведь для этого нужно не только разрабатывать новые алгоритмы, но и улучшать многие из ранее созданных.

GOOGLE ПРОТИВ PROMT

Кто переводит лучше?

Для тестирования переводчиков был использован текст новости BBC (раздел Technology). Результаты перевода приведены в таблице 1.

Подведем итоги нашего теста. Перевод с помощью Google Translate пока больше напоминает лотерею, и если повезет, вы можете получить хороший перевод документа. Если не повезет и смысл документа будет искажен, разработчики предусмотрительно называют сервис бета-версией — какие могут быть претензии?

Переводчик PROMT 8.0 достойно справился с переводом всех текстов, предложенных для тестирования. В ряде случаев полученный вариант требует правки, но смысл исходного текста всегда передается верно. Наличие возможностей настройки делает систему универсальным инструментом для перевода. Поэтому мы рекомендуем полагаться на традиционную технологию и использовать проверенный PROMT, чтобы получить грамотный перевод.

Таблица 2

Исходный текст BBC	Перевод Google	Перевод PROMT 8.0
<p>Under Pennsylvania law, the Department of General services has exclusive right to purchase the insurance contracts. Consequently, no employee of the University should sign any agreement with an insurance company. Should insurance be required for students and/or student actions or projects, the President of the University shall request West Chester University Student Services, Inc. to acquire such insurance and all contracts for institutional insurance shall be administered by the Vice President of Administrative and Fiscal Affairs.</p>	<p>В Пенсильвании права, Департамент Генерального услуг имеет исключительное право на приобретение договоров страхования. Следовательно, сотрудник университета должен подписать любое соглашение со страховой компанией. Если страхование необходимо для студентов и / или студенческих мероприятий или проектов, Президент университета просит Запад Честер университета Student Services, Инк. "На приобретение такого страхования и все договоры страхования институциональных под управлением вице-президента по административным и налоговым вопросам.</p>	<p>Согласно закону Пенсильвании, Отдел Общих услуг имеет исключительное право купить страховые контракты. Следовательно, никакой служащий Университета не должен подписать соглашение со страховой компанией. Если страхование требуется для студентов и/или студенческих действий или проектов, президент Университета должен просить West Chester University Student Services, Inc приобрести такое страхование, и всеми контрактами для установленного страхования должен управлять Вице-президент Административных и Финансовых Дел.</p>

Тестируем переводчики на тексте контракта

Как ни странно, но даже при переводе текста типового контракта Google сумел наделать много ошибок. Самой грубой является искажение смысла документа. Так, при переводе фразы «Consequently, no employee of the University should sign any agreement with an insurance company» Google полностью изменил смысл на противоположный. Вместо правильного перевода («никто из сотрудников не должен заключать соглашение со страховой компанией») Google выдал «сотрудник университета должен подписать любое соглашение со страховой компанией». Также можно отметить, что система не распознала, что West Chester University Student Services, Inc. — это название организации, и перевела все слова из названия по отдельности.

В переводе PROMT никаких серьезных ошибок не обнаружено. Единственным минусом является использование обшей лексики вместо полагающихся в таких случаях более строгих формулировок и специальных терминов. Например, вместо «исключительное право купить страховые контракты» (перевод PROMT) лучше сказать «исключительное право на приобретение договоров страхования» (перевод Google Translate). Справедливости ради стоит отметить, что нам удалось найти пример удачного перевода типового документа (контракт на поставку) с помощью Google (см. таблицу 3). Перевод Google Translate практически идеален, что можно объяснить наличием в базе системы готового перевода.

Таблица 3

Исходный текст BBC	Перевод Google	Перевод PROMT 8.0
<p>Should any disagreement arise between the parties as it relates to employment, each party shall be responsible for attorneys' fees and court costs.</p>	<p>В случае возникновения разногласий между сторонами в том, что касается занятости, каждая из сторон несет ответственность за оплату услуг адвокатов и судебные издержки.</p>	<p>Если любое разногласие возникает между сторонами, поскольку оно имеет отношение с занятостью, каждая сторона должна быть ответственной за судебные издержки и судебные издержки.</p>

Удачный пример подстановки Google готового перевода

Этот результат перевода наводит на мысль о том, что разработчикам лингвистических переводчиков нужно задуматься о возможной интеграции обеих технологий для получения оптимального качества перевода. Зачем каждый раз заново переводить типовые выра-

жения, если их можно один раз занести в базу и использовать многократно? Очевидно, что в этом случае потребуется создавать достаточно объемные базы параллельных текстов, но при этом и удовлетворенность пользователей от продукта будет выше.