

Теоретический материал

Системы перевода и распознавания текстов

В современном мире происходит очень важный процесс — формирование единого информационного пространства. Стираются информационные границы между странами и народами, у человека появляется возможность общаться в буквальном смысле слова со всем миром. Все это приводит к тому, что многие люди различных профессий начинают общаться с иностранными коллегами, читать справочную и другую специальную литературу на иностранном языке. Но далеко не каждый человек свободно владеет иностранными языками.

Современные компьютеры способны хранить большие массивы данных и производить в них быстрый поиск. Эти возможности компьютера можно использовать для создания электронных словарей и организации с их помощью перевода текста с одного языка на другой. Для этих целей сегодня уже существует множество программ.

Как работают программы-переводчики

Чтобы найти перевод неизвестного иностранного слова, пользователю электронного словаря достаточно ввести это слово в строке поиска, и уже через несколько мгновений будет получен исчерпывающий перевод. Современные текстовые процессоры имеют в своем составе словари, позволяющие производить орфографическую проверку правильности написания слов (на разных языках).

Но перевод отдельного слова и перевод целого текста — задачи совершенно разные. Чтобы понять смысл текста, не всегда хватает понимания значений всех входящих в него слов. Например, в английском языке слово «unit» имеет как минимум 6 различных значений. Какое из них имел в виду автор конкретного текста? Следствием необходимости решения этих проблем стало появление компьютерных систем перевода текстов. Современные системы перевода позволяют не только переводить, но и редактировать перевод, работать с различными тематическими словарями, выполнять как простой и быстрый, так и сложный и профессиональный перевод. Эти программы (вернее, пакеты программ) позволяют работать с файлами различных типов, электронной почтой, гипертекстовыми документами и т. п. К сожалению, задача адекватного перевода до конца еще не решена — многие программы зачастую выполняют ее не всегда удачно.

Рассмотрим простой пример. Переведем с помощью системы перевода на английский язык фразу:

Информатика — это наука об информации.

Результат перевода:

The computer science is an information science.

А теперь с помощью той же программы переведем эту фразу на русский язык. Получим:

Информатика — информатика.

Как говорится, почувствуйте разницу!

Системы перевода еще уступают человеку, особенно в работе с художественными текстами, но эта область информатики развивается очень быстро и «электронные карманные переводчики» уже становятся незаменимым помощником туриста, отправляющегося в страну с незнакомым для него языком.

Словари и переводчики

Словарь - это неотъемлемая часть библиотеки каждого интеллигентного человека. Ими также пользуются самые любопытные дети. А учителя и преподаватели утверждают - гортание словаря "от нечего делать" очень полезное дело.

Без них не обойдется ни студент, ни школьник, ни даже ученый. Специалисты из разнообразных сфер часто заглядывают в словари по тысячам вопросам. Даже очень хорошие переводчики перед работой, все равно кладут на стол несколько словарей.

Словари и электронная книга оказались очень взаимосвязанными между собой - за последнее время компьютерный словарь научился самостоятельно искать, находить и озвучивать искомое слово. Иногда он будет полезен и другим членам семьи при переводе электронного письма от знакомого из-за рубежа, или каких-либо инструкций к технике.

Беспрецедентным преимуществом такого рода словарей есть его цена и удобность: они помещаются на практически любом носителе от компакт-диска до "флешки".

Итак, на сегодня, для домашнего применения можно выделить несколько компьютерных словарей: *ABBYU Lingvo*, *Pragma*, *SlovoEd*, *Magic Gooddy*, *"Сократ Персональный"* и т.д.

Основные параметры машинных переводчиков

Параметры машинных переводчиков должны удовлетворять четырем основным требованиям:

- оперативность
- гибкость
- скорость
- точность

Оперативность заключается в возможности постоянного обновления словарного запаса и тематических разделов.

Гибкость рассчитана на конкретную предметную область.

Скорость - возможность автоввода и обработки текстовой информации с бумаги. Одна такая система (OCR-System) ежедневно заменяет больше десяти опытных машинисток.

Точность заключается грамотности и адекватной передачи смысла переводимого текста на язык перевода.

Языки перевода

Онлайн переводчик текстов поддерживает направления перевода для таких языков:

- английский
- немецкий
- русский
- французский
- и многие др.

Ввод текста и выбор направления перевода

Исходный текст нужно напечатать либо скопировать в верхнее окно и выбрать направление перевода из выпадающего меню. Например, для **русско-украинского перевода**, нужно ввести текст на русском языке в верхнее окно и выбрать из выпадающего меню пункт **«русский»**, затем **«украинский»**. Далее необходимо нажать клавишу **Перевести**

Специализированные словари

Если исходный текст для перевода относится к специфической отрасли, выберите **тему** специализированного лексического словаря из ниспадающего списка, например, Бизнес, Интернет, Законы, Музыка и другие. По умолчанию используется словарь общей лексики.

Проверка орфографии

Качество перевода зависит от правильности написания исходного текста. Советуем воспользоваться Проверкой орфографии. **Проверка орфографии** работает для украинского, русского и английского языков.

Транслитерация

При переписке с адресатом, у которого не установлена кириллица, можно воспользоваться **транслитерацией**. Транслитерация поддерживает русский и украинский языки, и транслитерирует как с латиницы в кириллицу, так и с кириллицы в латиницу.

Виртуальная клавиатура

Если необходимой раскладки нет на Вашем компьютере, воспользуйтесь **виртуальной клавиатурой**. Виртуальная клавиатура предлагается для русского, украинского, английского, немецкого, французского, испанского и итальянского языков.

Улучшение качества перевода

Существуют способы улучшения результатов машинного перевода:

1. Перед началом перевода, нужно определить тип текста, то есть из какой области жизнедеятельности человека он представлен (экономика, спорт, наука и т.д.). Ведь каждая сфера имеет свои нюансы и термины.

2. Часто причиной неправильного перевода являются опечатки переводимом тексте. Это касается и распознанных текстов. Слова с

ошибками помечаются переводчиками как незнакомые, потому что в таком виде их нет в словарях. Хуже, если есть ошибки в пунктуации - одна неправильно поставленная запятая способна исказить перевод всего предложения.

3. Работайте с фрагментами текста. Никогда не переводите весь текст сразу. В нем всегда найдутся слова, отсутствующих в словаре и такие, которые система переводит неправильно.

Обзор некоторых программ переводчиков

Информация в Сети представлена не только на русском языке. Русскоязычный сегмент Интернета – это только лишь небольшая часть того, что можно найти в Глобальной Сети. Рано или поздно вам придется столкнуться с сайтами на других языках, например, на английском. Если у вас возникнут затруднения при понимании иностранных слов, вам помогут программы-переводчики. Традиционно этот сегмент программного обеспечения относится к бизнес-классу – большинство программ такого типа платные и стоят довольно дорого. Мы рассмотрим несколько более бюджетных решений, достаточно интересных.

PROMT Express. Эта программа – один из лучших переводчиков из известных на сегодняшний день. Базовая бесплатная версия программы, работоспособная в течение 30 дней, умеет работать с английским языком – переводить с английского на русский и наоборот.

Чтобы перевести слово, можно ввести его в окне программы PromtX и нажать кнопку Перевести на панели инструментов.

Кроме того, программа умеет отслеживать содержимое буфера обмена – фрагмент текста, скопированный в буфер, автоматически будет помещен в окно программы-переводчика.

Сохранить результат перевода можно в RTF- или TXT-файле.

Если вы хотите получить синхронный перевод, который будет отображаться в окне перевода по мере ввода текста, выберите Перевод > Синхронный перевод.

Если в вашей программе подключено несколько словарей, вы можете изменить базовый словарь, выбрав Тематика > Словари документа. Если вы хотите зарезервировать некоторые слова для того, чтобы оставлять их без перевода, выберите Тематика > Зарезервированные слова. Слова, которые добавлены в этот список, не будут переводиться с помощью программы. В такой список стоит добавить различные имена собственные, возможно, непереводимые названия. А пометив переключатель Транслитерировать, вы заставите программу выводить зарезервированное слово латинскими буквами.

Web Translator. Эта программа умеет переводить сразу в нескольких направлениях. Среди языков, с которыми умеет работать Web Translator – английский, французский, испанский, немецкий, португальский, итальянский, русский и еще несколько других.

Программа умеет работать в двух режимах – в режиме Web и в режиме Text. Переключает их одноименная кнопка на панели инструментов программы.

В режиме Text программа позволяет ввести текст с клавиатуры или же вставить его из буфера обмена. При этом доступны инструменты форматирования текста, а также работа с буфером обмена.

Кроме того, программа может переводить содержимое RTF- и TXT-файлов. Для начала вам нужно открыть файл, выбрав File > Open Source Document. Поменять местами значения языков оригинала и перевода вы можете с помощью кнопки Swap.

На панели инструментов, выбирая нужные значения из списка, установите языки оригинала и перевода.

VuDictionary. VuDictionary – это простой словарь для перевода английских слов, который умеет работать в фоновом режиме. Программа может переводить только с английского на русский, работая со стандартным словарем.

Работает она так: пользователь вводит слово в строке, по ходу чего из словаря отбираются подходящие варианты перевода. Комплект отобранных вариантов отображается в виде списка, при этом отображается только ближайший перевод. Полный набор вариантов можно посмотреть в специальном окне, выделив нужное слово.

Слова в окне программы размещены по алфавиту. После удачного перевода найденное слово запоминается в списке истории, и в будущем перевод можно просмотреть, не прибегая к поиску.

Кстати, окно с переводом содержит очень детальную информацию, включая транскрипцию, варианты словоупотребления, варианты перевода.

TranslatIt! Эта программа отображает перевод во всплывающем окне рядом с незнакомым словом на web-странице. Для этого достаточно навести на него мышиную стрелку.

После запуска приложение размещает свой значок в системном трее, откуда и можно управлять его работой: изменить направление перевода, задать настройки, проверить обновления. Если программа работает в активном режиме (о чем сигнализирует пиктограмма, расположенная в трее), то для перевода незнакомого слова в браузере нужно просто задержать мышку над словом. Через несколько мгновений перевод появится рядом со словом.

Если постоянно появляющееся всплывающее окно с переводом слова мешает, можно установить режим отображения перевода только при нажатой клавише Ctrl.

Lingvo OnLine! Плагин Lingvo OnLine!, который позволяет переводить слова с английского на русский и обратно с помощью одного из двух сервисов – lingvo.yandex.ru или lingvo.ru.

В результате установки плагин интегрируется в контекстное меню. Для перевода слова вам необходимо выбрать Translate with Lingvo и указать, с какого языка необходимо получить перевод. В результате во всплывающем

окне будет отображен результат работы переводчика с переводом выделенного слова.

Онлайн переводчики: краткий обзор

Без знания хотя бы одного иностранного языка сегодня живется туго. Тем ценно, что в Интернете появилось бесчисленное множество программ-помощников, которые всегда готовы помочь во время чтения газеты или книги в оригинале.

Translate.ru (PROMT). Этот сервис от компании PROMT поддерживает порядка 19 языков и по праву занимает лидирующие позиции на рынке. Программа может переводить как отдельные слова, так и крылатые и устойчивые выражения, а также текст целиком. Получить более качественный перевод пользователь сможет после того, как укажет тематику.

Есть также интегрированный словарь, который показывает подробную грамматическую справку и примеры использования того или иного слова в контексте. Программу можно интегрировать на сайт в качестве виджета или кнопки для перевода сайта целиком.

Плюсы:

- можно усовершенствовать перевод, уточнив тематику;
- умеет переводить веб-страницы;
- содержит встроенные учебники и справочники по грамматике.

Минусы:

- ограничение в 3000 символов.

Google Translate. Сервис от специалистов компании Google способен переводить части или веб-страницы целиком на 103 языка. По количеству словарей и доступных функций, Google Переводчик - самый функциональный и универсальный сервис на сегодняшний день.

Если переводить отдельные слова, сервис автоматически переходит в режим онлайн-словаря, предлагая альтернативы с краткой характеристикой к каждому слову, показывает транскрипцию и транслитерацию, а также предоставляет озвучку.

Плюсы:

- большая языковая база;
- наличие транскрипции и транслитерации;
- поддержка функции озвучивания.

Минусы:

- ограничивает размер текста до 5000 символов;
- хромает перевод веб-страниц, чаще всего некорректный.

Яндекс Переводчик. Является одновременно словарем и сервисом для перевода больших текстов и веб-страниц. К каждому слову предлагается несколько вариантов значений слова на другом языке и подбор синонимов.

Полезная функция - программа умеет переводить текст с изображения, поддерживает как голосовой, так и текстовый ввод. Стоит

отметить и опцию предугадывания слов по смыслу, что существенно экономит время при вводе текста.

Также есть мобильное приложение, доступное на большинстве платформ, и поддержка более чем 40 языков. Мобильная версия может работать в офлайн-режиме, если дополнительно установить приложение программы.

Плюсы:

- интуитивные подсказки для текстового набора;
- предлагает синонимы и альтернативные варианты;
- офлайн-режим.

Минусы:

- есть языки, которым нужна доработка.

Мультитран. Ключевой особенностью сервиса является форум, где пользователи сервиса могут попросить о помощи и спросить совета.

Программа предоставляет детальный перевод, подбирает к словам список синонимов, поддерживает возможность как голосового, так и текстового ввода информации (плюс есть функция "прослушать").

Есть опция построчного перевода: если в строке программа обнаружила устойчивое выражение, оно дополнительно выделяется, а в сноске показывается пояснение.

Плюсы:

- форум, где можно попросить о помощи в переводе;
- обширная база для поиска синонимов;
- пояснение устойчивых выражений.

Минусы:

- пользователи отмечают нестабильность в работе сервиса в рабочее время;
- много ненужной информации.

Reverso. Сервис для перевода на 10+ языков мира в режиме онлайн. На выбор есть опция перевода на сайте программы, браузерное расширение и мобильная версия.

Программа подходит для несложных текстов, не содержащих специфическую терминологию, идиоматических и сленговых выражений, так как разработчики не гарантируют качественного перевода. Особенно это касается статей имеющих узкоспециализированную терминологию.

Переводить можно в нескольких режимах, в зависимости от поставленных пользователем задач. Например, есть обычный словарь с колонками для ввода текстов, раздел для слов и словосочетаний, а также проверка форм слова.

Плюсы:

- есть возможность проверки грамматики;
- переводит с подгруженных файлов и веб-страниц;
- можно отправлять перевод по почте.

Минусы:

- подходит исключительно для коротких и незамысловатых текстов.

Microsoft Translator. Сервис от компании Microsoft поддерживает 60 с лишним языков. Помимо браузерной версии есть мобильное приложение для Android, iOS, а также Apple Watch. В мобильной версии есть офлайн-режим, только необходима загрузка словарей.

Сервис переводит не только тексты, но и голосовые сообщения, фотографии и скриншоты. Для каждого слова программа предлагает альтернативы.

Программа также предлагает пользователю специальные разговорники и различные руководства.

Плюсы:

- большая база поддерживаемых языков;
- функция синхрона;
- есть функция голосового перевода;
- распознает текст на фотографиях;
- есть синхронизация на разных приложениях.

Минусы:

- функция перевода с картинки или скриншота недоработана

SYSTRANet. Так называемый сервис-старожил, который более 40 лет предлагает свои услуги на рынке онлайн. Может работать на различных платформах: от стационарных версий для персональных компьютеров до серверов. Предоставляет услуги перевода на более чем 130 языков, а благодаря широким возможностям по умолчанию интегрирован на устройствах серии Samsung Galaxy S и Note.

Сервис способен к самообучению, что позволяет пользователю максимально его кастомизировать под себя. Может переводить тексты, веб-страницы и загруженные файлы (txt, htm, rtf). Для качественного перевода можно выбрать тематический словарь или же создать пользовательский.

Плюсы:

- многоплатформенный сервис (работает на Windows, MacOS, Android, iOS);
- способен к самообучению;
- переводит не только тексты, но и файлы, интернет-страницы;
- есть тематические словари;
- есть функция перевода RSS-лент.

Минусы:

- ограничен размер текста до 1000 слов;
- информация о ресурсе есть только на английском языке.

Free Translation. Сервис для перевода рукописных текстов, загруженных документов, веб-страниц на 80+ языков. Сервис имеет ряд полезных опций, правда пользователю они станут доступны при покупке платного абонемента.

В случае, если пользователю недостаточно полученного онлайн-перевода, можно заказать услугу у квалифицированных специалистов.

Плюсы:

- умеет переводить веб-страницы целиком;
- услуга заказного перевода от специалистов.

Минусы:

- часть опций возможны при покупке платной версии.

Worldlingo. Онлайн-транслейтер, база которого охватывает 32 языка. Его "миссия" - переводить тексты, загруженные документы, правда, все это с ограничением до 1000 символов. Так же есть ограничение на перевод веб-сайтов, электронных сообщений, но на загруженные файлы оно не распространяется (все ограничения можно снять, купив платный тариф).

Этот сервис умеет переводить электронные письма, нужно будет лишь указать отправителя и адресата, выбрать тематику текста и ввести его, после чего письмо само отправится.

Плюсы:

- большая языковая база;
- можно выбрать тематику сайта/текста;
- может переводить электронные письма.

Минусы:

- только англоязычный интерфейс;
- не всегда точный перевод;
- медленная скорость загрузки страниц.

ABBYY Lingvo Live. Кроссплатформенный социальный сервис с открытым доступом к онлайн - словарям.

Благодаря ABBYY Lingvo юзер может самостоятельно найти перевод для каждого слова или выражения. Сервис также показывает форму слова, возможные альтернативные значения, транскрипцию; можно прослушать произношение и примеры лаконичного употребления того или иного слова.

Для удобства можно пользоваться экранной клавиатурой. Основная фишка - мощное интернет-сообщество, к которому можно обратиться за переводом редкого или узкоспециализированного слова, но для этого нужно быть зарегистрированным пользователем.

Плюсы:

- интернет-сообщество, которое заменит машинный перевод;
- удобный интерфейс, ничего лишнего.

Минусы:

- нельзя перевести большой текст, файлы тд.;
- находится на стадии бета-разработки.

Распознавание текста

Перед обсуждением этой темы давайте вспомним, какие устройства ввода информации существуют у современных компьютеров? Клавиатура, мышь, сканер и др. Сканер, например, позволяет вводить графическую информацию с листа бумаги.

За сотни лет человечество накопило огромный объем информации на традиционных бумажных носителях (книгах, газетах, журналах и т. п.), В настоящее время существует потребность (у электронных библиотек, к примеру) переносить эту информацию в память компьютера. Конечно, это можно сделать с помощью клавиатуры и текстового редактора, но, представьте себе, сколько времени уйдет даже у профессионального оператора на ввод, скажем, романа «Война и мир»? Необходимо как-то ускорить этот процесс. Встает вопрос, нельзя ли использовать сканер для ввода текстовой информации? Правда, в этом случае возникает такая проблема: все, что введено с помощью сканера, хранится в памяти ЭВМ как изображение. Надо «объяснить» компьютеру, что значок«с» — не просто закорючка, а буква, и хранить и обрабатывать его нужно как букву,

Ввод в компьютер печатного и рукописного текста

Существуют программы, позволяющие вводить тексты в ПК с помощью сканера. Используя специальные алгоритмы, они распознают буквы, позволяют редактировать распознанный текст и сохранять его в различных форматах. Популярной программой такого типа является АBBY FineReader. Работать с этой программой несложно. Сначала нужно отсканировать текст (управлять сканером можно прямо в среде FineReader), затем разбить этот текст на фрагменты, потом распознать эти фрагменты, отредактировать полученный текст и, наконец, сохранить его в нужном текстовом формате. Интерфейс программы позволяет освоить эти операции легко и быстро.

Задача распознавания текста относится к области проблем, которые решает наука под названием «Искусственный интеллект». Современные распознающие программы умеют читать не только печатный текст, но и текст, написанный самым «корявым» почерком.

Системы распознавания текста

С помощью сканера достаточно просто получить изображение страницы текста в графическом файле. Однако работать с таким текстом невозможно: как любое сканированное изображение, страница с текстом представляет собой **графический файл** - обычную картинку. Текст можно будет читать и распечатывать, но **нельзя его редактировать и форматировать**. Для получения документа в формате текстового файла необходимо провести **распознавание текста**, то есть преобразовать элементы графического изображения в последовательности текстовых символов.

Преобразованием графического изображения в текст занимаются специальные программы распознавания текста (Optical Character Recognition - OCR). Современная OCR должна уметь **распознавать тексты**, набранные не только определенными шрифтами (именно так работали OCR первого

поколения), но и самыми экзотическими, вплоть до рукописных, распознавать не только четко набранные тексты, но и такие, качество которых, мягко говоря, далеко от идеала. Например, текст с пожелтевшей газетной вырезки или третьей машинописной копии. Само собой, распознать текст — это еще полдела. Не менее важно обеспечить возможность **сохранения результата в файле популярного текстового формата** — скажем, формата Microsoft Word.

Как видим, для того, чтобы получить электронную, готовую к редактированию копию любого печатного текста, программе OCR необходимо выполнить «цепочку» из множества отдельных операций:

Сначала необходимо распознать структуру размещения текста на странице: выделить колонки, таблицы, изображения и так далее. Далее выделенные текстовые фрагменты графического изображения страницы необходимо преобразовать в текст.

Если исходный документ имеет типографское качество (достаточно крупный шрифт, отсутствие плохо напечатанных символов или исправлений), то задача распознавания решается методом сравнения с растровым шаблоном. Сначала растровое изображение страницы разделяется на изображения отдельных символов. Затем каждый из них последовательно накладывается на шаблоны символов, имеющихся в памяти системы, и выбирается шаблон с наименьшим количеством отличных от входного изображения точек. При распознавании документов с низким качеством печати (машинописный текст, факс и так далее) используется метод распознавания символов по наличию в них определенных структурных элементов (отрезков, колец, дуг и др.).

Любой символ можно описать через набор значений параметров, определяющих взаимное расположение его элементов. Например, буква «Н» и буква «И» состоят из трех отрезков, два из которых расположены параллельно друг другу, а третий соединяет эти отрезки. Различие между данными буквами — в величине углов, которые образует третий отрезок с двумя другими.

При распознавании структурным методом в искаженном символьном изображении выделяются характерные детали и сравниваются со структурными шаблонами символов. В результате выбирается тот символ, для которого совокупность всех структурных элементов и их расположение больше всего соответствует распознаваемому символу.

Наиболее распространенные системы оптического распознавания символов, например, **ABBYY FineReader** и **CuneiForm** от Cognitive, используют как растровый, так и структурный методы распознавания. Кроме того, эти системы являются «самообучающимися» (для каждого конкретного документа они создают соответствующий набор шаблонов символов) и поэтому скорость и качество распознавания многостраничного документа постепенно возрастают.

Качество распознавания во многом зависит от того, насколько хорошее изображение получено при сканировании. Качество изображения

регулируется установкой основных параметров сканирования: типа изображения, разрешения и яркости.

Сканирование в сером является оптимальным режимом для системы распознавания. В случае сканирования в сером режиме осуществляется автоматический подбор яркости. Если Вы хотите, чтобы содержащиеся в документе цветные элементы (картинки, цвет букв и фона) были переданы в электронный документ с сохранением цвета, необходимо выбрать цветной тип изображения. В других случаях используйте серый тип изображения.

Оптимальным разрешением для обычных текстов является - 300 dpi и 400-600 dpi для текстов, набранных мелким шрифтом (9 и менее пунктов).

При заполнении налоговых деклараций, при проведении переписей населения и так далее используются различного вида бланки с полями. Рукопечатные тексты (данные вводятся в поля печатными буквами от руки) распознаются с помощью систем оптического распознавания форм и вносятся в компьютерные базы данных.

Сложность состоит в том, что необходимо распознавать написанные от руки символы, довольно сильно различающиеся у разных людей. Кроме того, система должна определить, к какому полю относится распознаваемый текст.

Системы распознавания рукописного текста. С появлением первого карманного компьютера Newton фирмы Apple в 1990 году начали создаваться системы распознавания рукописного текста. Такие системы преобразуют текст, написанный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.