Гибридная технология перевода

Юлия Епифанцева
PROMT

PROMT®

Машинный перевод

Машинный (автоматический) перевод – процесс перевода текстов с одного естественного языка на другой с помощью компьютерной программы

Основные типы систем МП

- Rule-based машинный перевод (RBMT) – перевод, основанный на правилах.
- Статистический машинный перевод (SMT).

➔ Гибридные системы перевода (HMT)

Типы RBMT

- **Системы по типу Transfer**

  предложение на языке входа =>
  морфологический, грамматический, семантический анализ =>
  преобразование в структуру выходного языка (TRANSFER) =>
  синтез выходного предложения по полученной структуре=>
  предложение на языке выхода

- **Системы по типу Interlingua**

  предложение на языке входа =>
  анализ входного предложения в терминах метаязыка =>
  синтез из метаструктуры предложения выходного языка =>
  предложение на языке выхода

  Разработка метаязыка = языконезависимое представление, наличие знаний о мире (онтологии, логики предикатов)

Компоненты RBMT на примере PROMT

- Лингвистические базы данных
  - двуязычные словари
  - файлы имен, транслитерации
  - морфологические таблицы
- Модуль перевода
  - грамматические правила
  - алгоритмы перевода

Двуязычные словари

имеют трехуровневую структуру для настройки системы на различные предметные области:

- **Генеральный** словарь (от 50 до 250 тысяч статей)
- **Специализированные** словари (от 5 до 100 тысяч статей; охватывают различные тематики: бизнес, спорт, IT, добыча нефти и газа, металлургия...)
- **Пользовательские** словари (вспомогательные, открыты для редактирования пользователю)

Словарная статья (1)

- Слово или выражение на входном языке
- Коллекция переводов (активных и неактивных)
- Структурированное описание различной лингвистической информации (морфологической, синтаксической, семантической) в виде набора признаков и модификаторов

Словарная статья (2)

Словарная статья (3)

Словарная настройка на примере ‘chest’

Модификаторы	Перевод
ſXj	груд/n13/1
ſX_H8	комод/n1/1
ſPpb▪Yj	груд/n13/1
ſY_H8▪Amn	комод/n1/1
ſPiцю_la_lF_lE_lG▪	фонд/n1/1
ſVhOyv▪Yr	упаков/n4/1
FY	грудной клетки/a98/1
F¤	грудн/a5/1
ſPtv▪Y_H8	комод/n1/1
ſP_Rn▪Y_H8	комод/n1/1

Контекст	Перевод
This is a chest	Это грудь (комод)
There is a tattoo on his chest	На его груди есть татуировка
At the time of the chest examination the blood pressure may be taken	Во время обследования грудной клетки может быть измерено давление
The University Chest is a term used at Oxford in connection with the financial aspects of the university and its administration	Университетский фонд – термин, использованный в Оксфорде в связи с финансовыми аспектами университета и его администрации
The oak chest with iron lock	Комод из дуба с железным замком

Уровни анализа предложения (1)

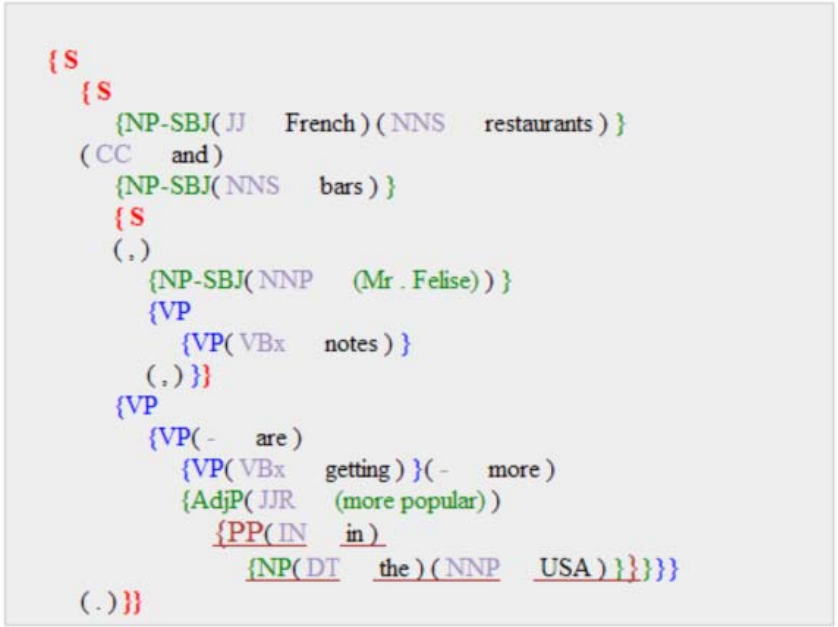
- Препроцессор
- Нормализация текста (удаление повторяющихся пробелов...)
- Токенизация входной цепочки (поиск входных словоформ в словаре с сопутствующим морфологическим анализом)
- Лексический анализ (контекстный анализ, различные склейки: имена, номера телефонов, даты...)
- Снятие омонимии (определение частей речи в случаях грамматической неоднозначности)
- Уровень сбора групп (соединение лексических единиц в группы)
- Анализ сложных предложений (выделение простых в составе сложного, синтаксическая омонимия)
- Семантико-синтаксический разбор (заполнение глагольного фрейма)
- Синтез (синтез по полученной структуре, расстановка элементов внутри группы и групп в предложении...)

Уровни анализа предложения (2)

French restaurants and bars, Mr. Felise notes, are getting more popular in the USA.

- Лексический анализ
  French restaurants and bars , [Mr. Felise] notes , are getting [more popular] in the USA .
- Синтаксические группы
  [French restaurants][and][bars] [,][Mr. Felise][notes] [,][are getting][more popular][in][the USA][.]
- Анализ сложного предложения
  { [**French restaurants**] [and][bars] { [,][**Mr. Felise**][notes] [,]} [are getting][more popular][in][the USA][.]}
- Анализ простых предложений
  { [**French restaurants**] [and][**bars**] { [,][**Mr. Felise**][notes] [,]} [are getting][more popular][in][the USA][.]}

Уровни анализа предложения (3)



Преимущества и недостатки RBMT

- Преимущества
  - синтаксическая и морфологическая точность,
  - стабильность и предсказуемость результата,
  - возможность настройки на предметную область.

- Недостатки
  - трудоемкость и длительность разработки,
  - необходимость поддерживать и актуализировать лингвистические БД,
  - «машинный акцент» при переводе.

Статистический машинный перевод (1)

История

- Принципы SMT разработаны еще в 1949 г. Уорреном Уивером
- «Вторая волна» – начало 1990-х, IBM
- «Третья волна» – Google, Microsoft, Language Weaver, Яндекс и десятки других

Статистический МП – это поиск наиболее вероятного перевода предложения с использованием данных, полученных из параллельных корпусов.

Статистический машинный перевод (2)

- Сегодня SMT – наиболее активно разрабатываемая архитектура MT. Почему?
  - Легко построить, если есть двуязычный корпус, ноль / минимум лингвистики
  - Переносимость технологии на любые пары языков
  - Лексическая гладкость
- Ограничения / недостатки:
  - Ограниченность параллельных корпусов в природе и их real-life качество
  - Плохо справляется с морфологией / синтаксисом (по сравнению с RBMT)
  - Искажение информации (дублирование, пропуск или подмена информации)

USA is to blame = США не виноват
Russia is to blame = Россия виновата

Выводы

Обе технологии имеют свои достоинства и недостатки, но главное – они не решили задачу по получению качественного машинного перевода.

MT-сообщество ожидает прорыва в качестве перевода в гибридных моделях RBMT + SMT.

Гибридная технология PROMT

Объединение RBMT и статистических технологий

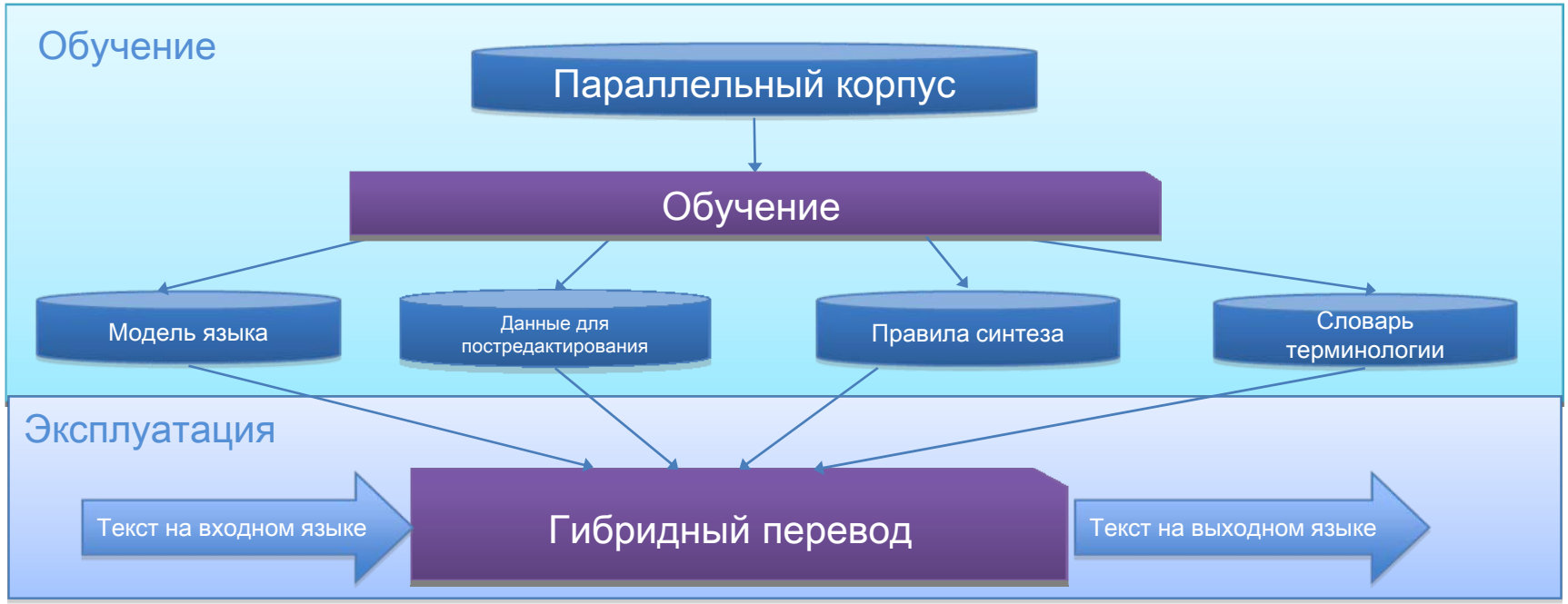
- лингвистический анализ входного предложения
- порождение вариантов перевода
- использование статистических технологий
- оценка и выбор лучшего варианта перевода с использованием Модели языка

Этапы Гибридной технологии

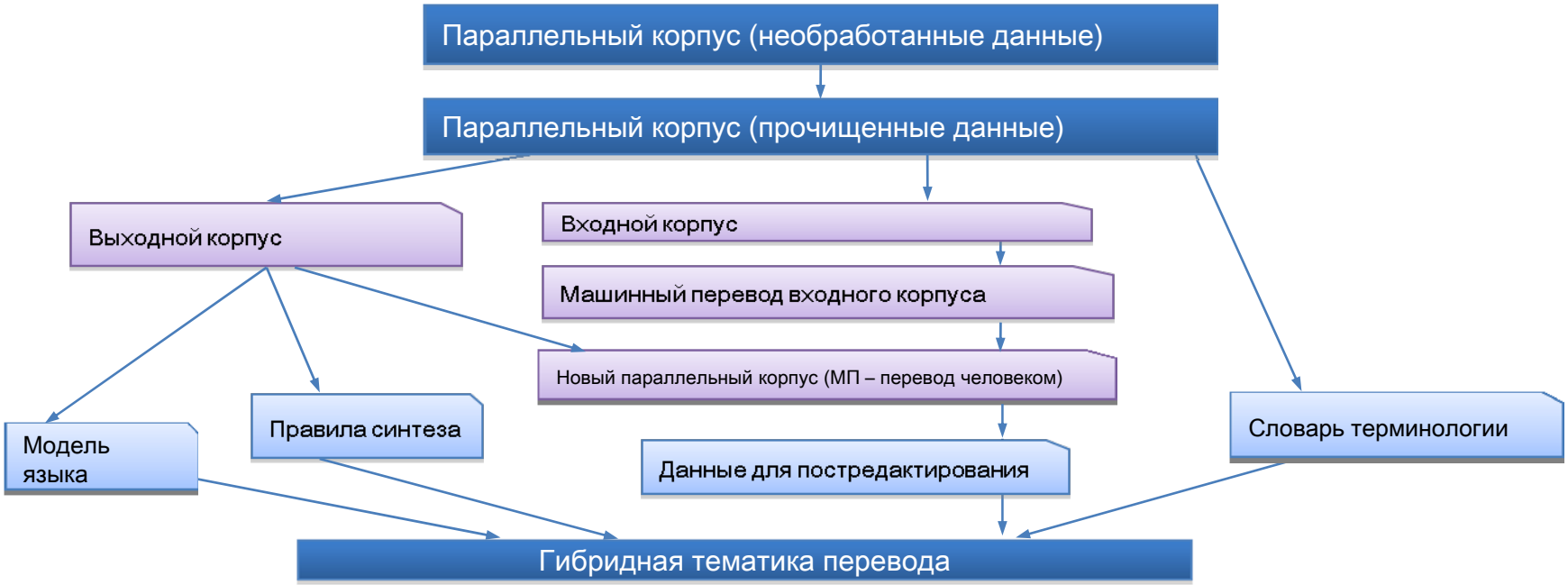
- Обучение RBMT на основе параллельного корпуса с использованием статистических технологий

- Эксплуатация на основе натренированной системы

Архитектура Гибридной технологии

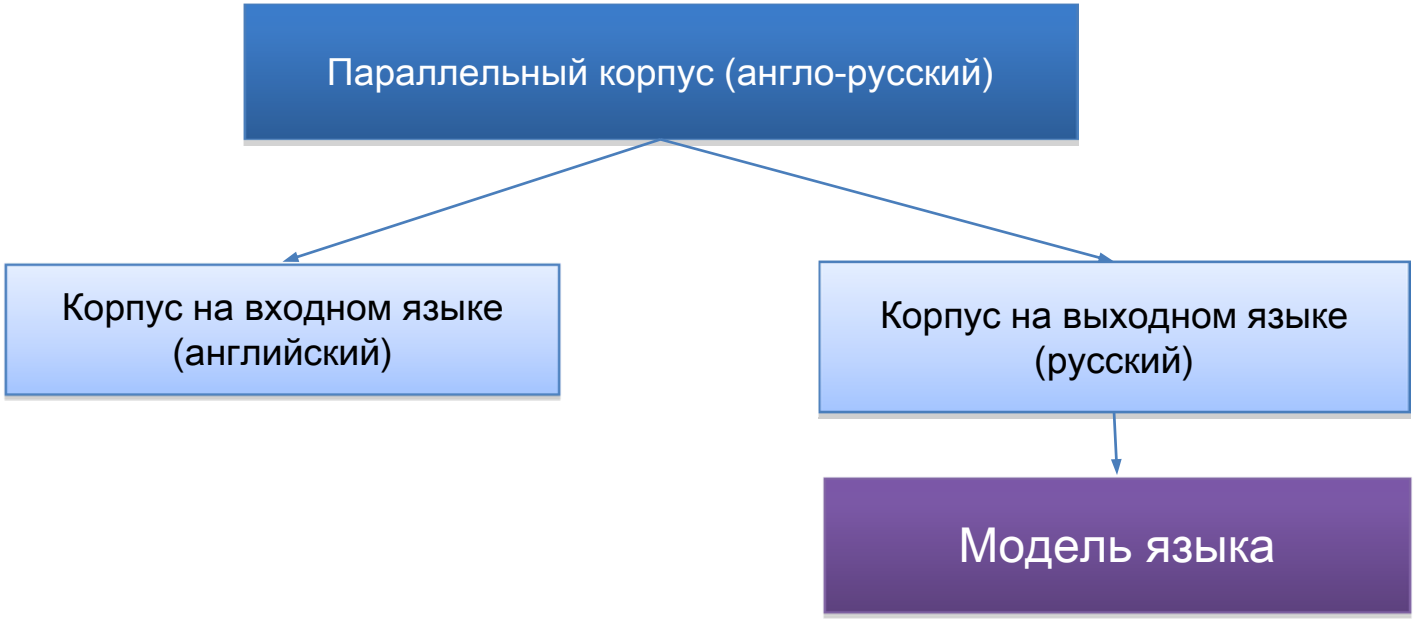


Обучение


Параллельный корпус (необработанные данные)
Параллельный корпус (прочищенные данные)
Выходной корпус
Входной корпус
Машинный перевод входного корпуса
Новый параллельный корпус (МП – перевод человеком)
Модель языка
Правила синтеза
Данные для постредактирования
Словарь терминологии
Гибридная тематика перевода

Эксплуатация

Модель языка (1)



Модель языка (2)

- Модель языка (Language Model/LM) – это набор n-грамм моноязычного корпуса с их вероятностными характеристиками.

- N-грамма – это последовательность слов из предложений длины n.

Модель языка (3)

```
-1.755048       по двум очередям в      -0.6624343
-1.755048       не к очередям каждого   -0.0527804
-1.755048       относящимся к очередям относятся        -0.05122143
-1.755048       обслуживания различным очередям и       -0.05385702
-1.056816       обеими входящими очередями и    -0.6624451
-0.4883884      между двумя очередями </s>
-1.056816       совместно используется очередями и      -0.6624537
-2.103502       буфера между очередями путем    -0.05193719
-2.103502       между различными очередями процент      -0.04642604
-2.188747       вызовы в очередях по    -0.05341507
-1.317873       использовать в очередях виртуальных     -0.5909631
-1.044643       находящимся в очередях уровня   -0.05364625
-1.044643       пакеты в очередях уровня        -0.05345588
-2.188747       потерями в очередях входа       -0.6321825
-2.118226       Во входящих очередях SRR        -0.6562098
-2.118226       и выходной очередях количество  -0.0537526
-0.435618       в других очередях </s>
-1.55669        особенно при очередях с -0.05344867
-1.641104       снятия изоляции очистите от     -0.662367
-1.641104       головного узла очистите статистику      -0.05261082
-0.8563652      <s> Чтобы очистить базу 0.1404502
-2.303862       поздних версиях очистить VLAN   -0.1768938
-2.303862       увеличится если очистить место  -0.1765835
-0.435618       NAT можно очистить с    -0.6248463
-1.742327       Если необходимо очистить сеанс  -0.6620195
-2.303862       если необходимо очистить все    -0.05370406
```

Модель языка (4)

- Perplexity (PPL) – вычисляемая для предложения «степень его искаженности» по отношению к данной LM. Чем меньше PPL, тем «естественнее» предложение.

- Модель языка
  - оценка релевантности (через PPL) каждого перевода по отношению к данному корпусу,
  - выбор лучшего варианта среди всех порожденных.

Как работает Гибридная технология

- Создание терминологического словаря из параллельных текстов для RBMT автоматическим путем
- Порождение всех возможных вариантов перевода на основе
  - лексических вариантов
  - вариантов синтеза разных конструкций
  - применения постредактирования

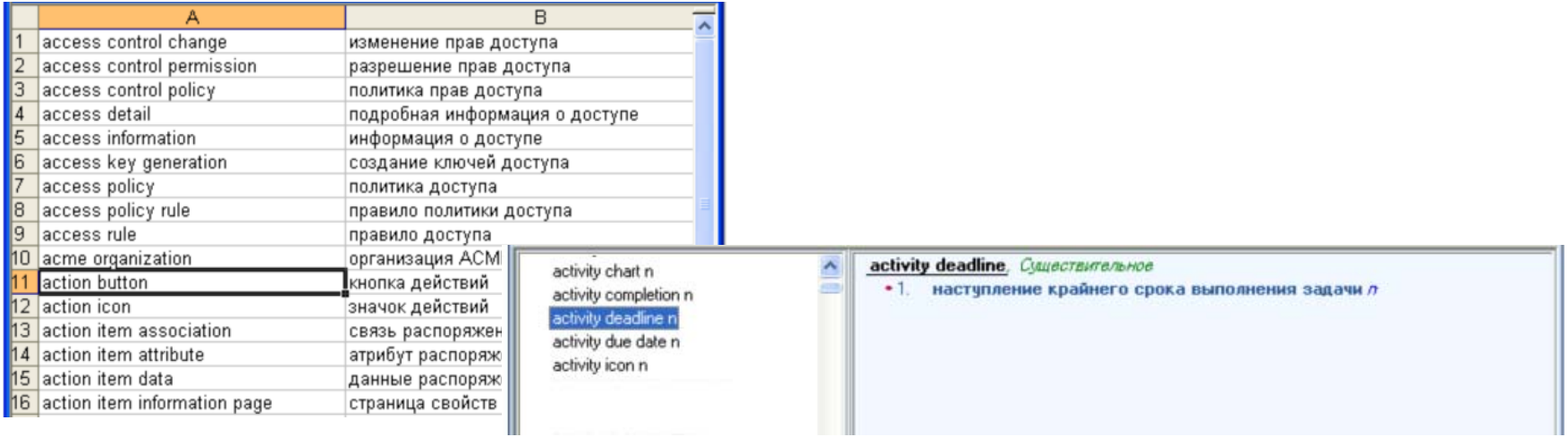
➔ выбор лучшего варианта через Модель языка

Терминологический словарь (1)

Технология получения:
а) на основе параллельного корпуса составляются таблицы n-грамм входного корпуса вместе с вероятностями соответствий этих n-грамм n-граммам выходного корпуса,
б) на основании парсеров для входного и выходного языков, а также частотных характеристик из общего числа n-грамм извлекаются релевантные для словаря единицы с некоторой грамматической информацией (например, о валентности)

➔ создается двуязычный глоссарий
в) в автоматическом режиме создается словарь для RBMT

Терминологический словарь (2)



	A	B
1	access control change	изменение прав доступа
2	access control permission	разрешение прав доступа
3	access control policy	политика прав доступа
4	access detail	подробная информация о доступе
5	access information	информация о доступе
6	access key generation	создание ключей доступа
7	access policy	политика доступа
8	access policy rule	правило политики доступа
9	access rule	правило доступа
10	acme organization	организация ACM
11	action button	кнопка действий
12	action icon	значок действий
13	action item association	связь распоряжен
14	action item attribute	атрибут распоряж
15	action item data	данные распоряж
16	action item information page	страница свойств

Лексические варианты

The restaurant features a unique space with a cozy lounge and an eclectic blend of music, art and sculpture.

Rule-based

Ресторан представляет собой уникальное пространство (место) с удобным залом и эклектичной смесью музыки, искусства и скульптуры.

Hybrid

Ресторан представляет собой уникальное пространство с удобным залом и эклектичной смесью музыки, искусства и скульптуры. (PPL==778)

Ресторан представляет собой уникальное место с удобным залом и эклектичной смесью музыки, искусства и скульптуры. (PPL=522)

Варианты синтеза конструкций (1)

Rule-based: выбор определенной модели синтеза
Hybrid: синтезирование нескольких вариантов перевода

Правило синтеза: перевод конструкции to + инфинитив
- чтобы + инфинитив
- для + существительное

You can use the same steps to edit the collection.

Можно использовать те же самые шаги, чтобы отредактировать коллекцию. *(PPL=372)*

Можно использовать те же самые шаги для редактирования коллекции. *(PPL=358)*

Варианты синтеза конструкций (2)

Rule-based: выбор определенной модели синтеза
Hybrid: синтезирование нескольких вариантов перевода

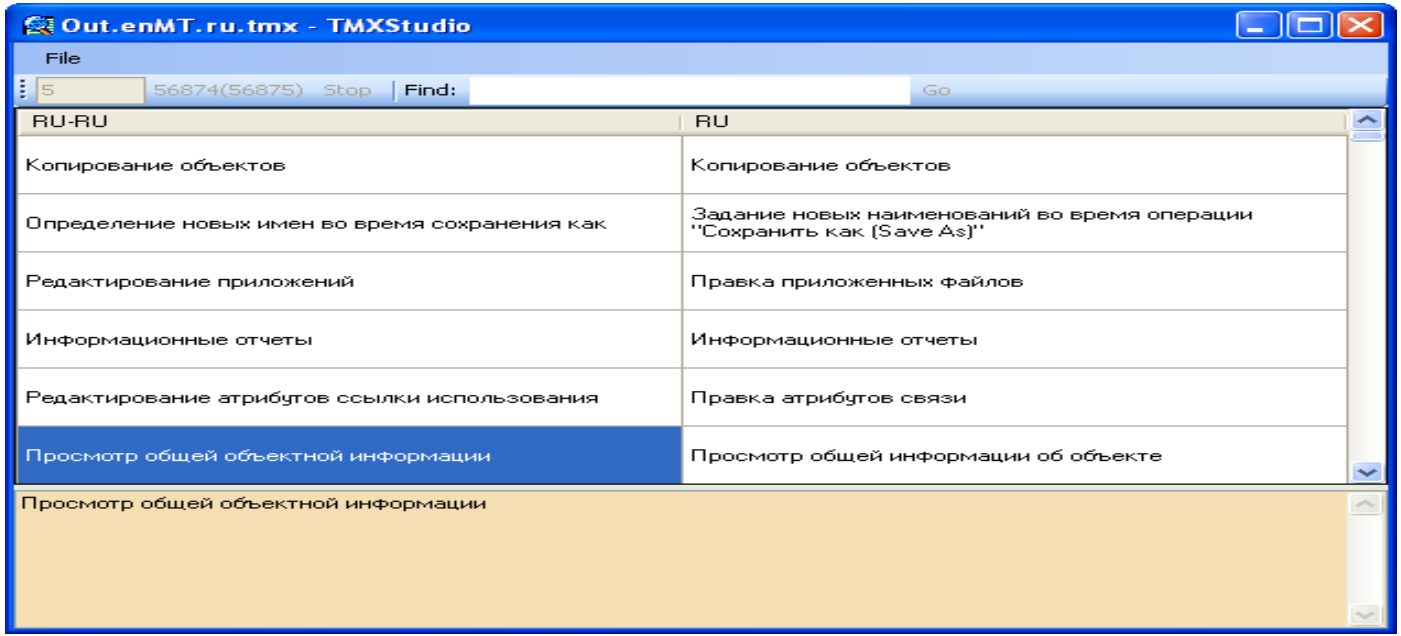
Правило синтеза: порядок следования подлежащего и сказуемого.

Click Browse to browse the path for the folder in which you want newly created documents to be located.

Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы недавно созданные документы были расположены. (PPL= 290)

Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы были расположены недавно созданные документы. (PPL= 274)

Данные для постредактирования (1)



Данные для постредактирования (2)

Технология : на основе параллельного корпуса выделяется таблица n-грамм входного корпуса вместе с вероятностями соответствий этих n-грамм n-граммам выходного корпуса.

- *с платежом PayPal банковским переводом → в случае платежа PayPal посредством банковского перевода*
- *вводите банковский перевод → инициируете перевод*
- *когда Вы закончены → после окончания Вашей работы*
- *каждое усилие было приложено → были предприняты все усилия*

Данные для постредактирования (3)

Пример применения нескольких замен сегментов машинного переводами сегментами человеческого перевода.

With PayPal payment by bank transfer, you initiate a bank transfer from your bank account to your PayPal account.

С платежом PayPal банковским переводом вы вводите банковский перевод с Вашего банковского счета на ваш счет PayPal. (PPL=95)

В случае платежа PayPal посредством банковского перевода вы инициируете перевод с Вашего банковского счета на ваш счет PayPal. (PPL == 7)


Исходный текст
Click Browse to browse the path for the folder in which you want newly created documents to be located.
Порождение лексических вариантов
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были расположены.
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были размещены.
Порождение вариантов синтеза
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были расположены.
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были размещены.
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы были расположены недавно созданные документы.
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы были размещены недавно созданные документы .
Порождение вариантов постредактирования
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были расположены.
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были размещены.
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы были расположены недавно созданные документы.
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы были размещены недавно созданные документы .
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы недавно созданные документы были расположены.
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы недавно созданные документы были размещены.
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы были расположены недавно созданные документы.
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы были размещены недавно созданные документы .

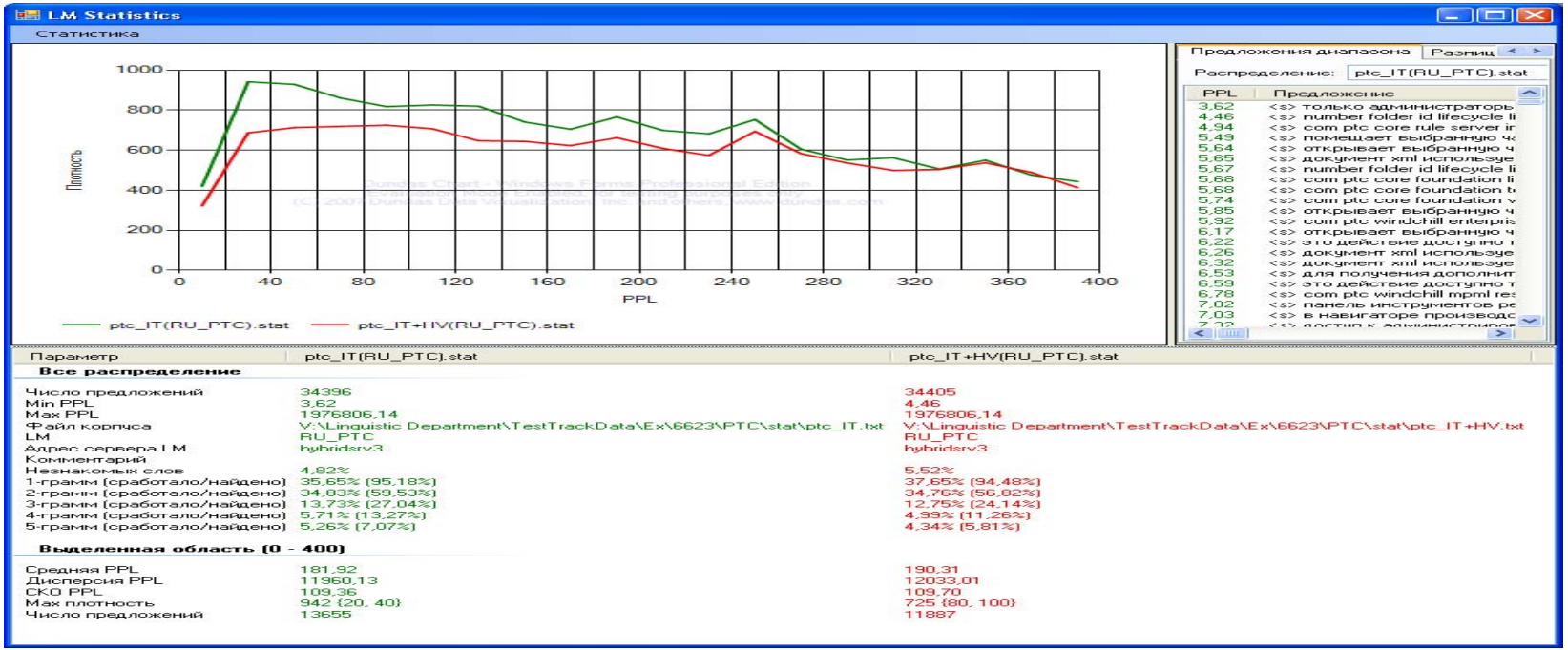
Оценка LM

Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были расположены. (PPL= 556)
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы недавно созданные документы были размещены. (PPl = 601)
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы были расположены недавно созданные документы. (PPl = 526)
Нажмите Browse, чтобы рассмотреть путь к папке, в которой Вы хотите, чтобы были размещены недавно созданные документы. (PPl = 569)
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы недавно созданные документы были расположены. (PPl = 277)
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы недавно созданные документы были размещены. (PPl = 301)
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы были расположены недавно созданные документы. (PPl = 261)
Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы были размещены недавно созданные документы. (PPl = 283)

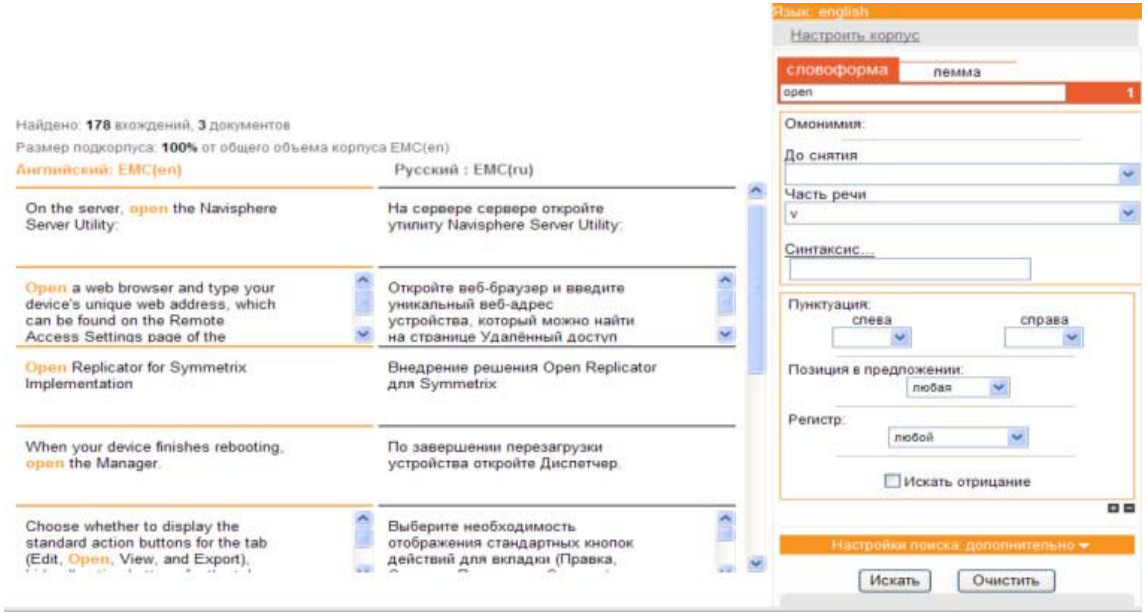
Выбор лучшего варианта

Нажмите Browse для просмотра пути к папке, в которой Вы хотите, чтобы были расположены недавно созданные документы. (PPl = 261)

LM Statistics



Поиск по параллельным корпусам



Выводы

- Преимущества RBMT сохраняются:
  - синтаксическая и морфологическая точность,
  - стабильность и предсказуемость результата,
  - возможность настройки на предметную область.

- Недостатки RBMT нивелируются за счет использования параллельных корпусов и статистических методов
  - автоматическая настройка лингвистических баз данных (быстрое и качественное извлечение терминологии),
  - исчезает «машинный» акцент при перевода (варианты синтеза и постредактирование).

Примеры перевода с помощью гибридной технологии

Исходный текст	Гибридный машинный перевод	Перевод человека
To replace any parts of devices or cables, use only original spare parts or parts which are explicitly licensed by the manufacturer.	Для замены любых частей устройств или кабелей используйте только первоначальные запасные части или части, явно лицензированные изготовителем.	Для замены любых компонентов устройств или кабелей должны использоваться только оригинальные запасные части или компоненты, явно лицензированные производителем.
Remove the faulty CF card from the SC card.	Удалите неисправную CF-карту из Платы SC.	Выньте неисправную CF-карту из платы SC.
It is a precondition for reuse and recycling of used electrical and electronic equipment.	Это является предварительным условие для вторичного использования и сброса используемого электрического и электронного оборудования..	Это является необходимым предварительным условием для вторичного использования и переработки отработавшего электрического и электронного оборудования.
Total Loss of Capacity Reception	Total Loss of Capacity Transmit; Полная потеря приема пропускной способности	Total Loss of Capacity Reception; Полная потеря пропускной способности при приеме
Document history	Хронология документации	Хронология документа
There is a risk of explosion if the battery is not replaced correctly!	Если батарея не заменяется правильно, существует риск взрыва!	При неправильной замене батареи существует опасность взрыва!
This document is intended for anyone tasked with installing, operating, maintaining, or troubleshooting the SURPASS hiT 7300 system.	Этот документ предназначен для специалистов, для которого определяют задачу с инсталляцией, эксплуатацией, обслуживанием, или устранению отказов SURPASS hiT 7300 систем.	Этот документ предназначен для персонала, отвечающего за процессы установки, эксплуатации, технического обслуживания или устранения отказов применительно к системе SURPASS hiT 7300.

Примеры перевода с помощью гибридной технологии

Исходный текст	Гибридный машинный перевод	Перевод человека
Unfortunately, if access to your account remains limited for an extended period of time, it may result in further limitations or eventual account closure.	К сожалению, если доступ к счету остается ограниченным в течение длительного периода времени, он может привести к дальнейшим ограничениям или конечному закрытию счета.	К сожалению, ограничение доступа к учетной записи на более длительный период времени может привести к дальнейшим ограничениям или к ее окончательному закрытию.
We thank you for your prompt attention to this matter.	Мы благодарим Вас за вашу оперативность по отношению к данному вопросу.	Благодарим вас за то, что вы сразу же обратили внимание на эту проблему.
Your account was temporarily limited as a security measure and full access has not been restored yet because we haven't heard from you.	Ваш счет был временно ограничен как меры безопасности, и полный доступ еще не был восстановлен, поскольку мы не получили известие от вас.	Ваш счет временно ограничен с целью защиты. Полный доступ не был восстановлен, потому что мы не получили от вас ответа.
We understand it's frustrating and inconvenient to have your account permanently limited but you'll still be able to see your transaction history for a limited time.	Мы понимаем ваше разочарование и неудобства, связанные с постоянным ограничением счета постоянно, но вы по-прежнему сможете просматривать историю трансакций.	Мы понимаем ваше разочарование и неудобства, связанные с постоянным ограничением счета, но некоторое время вы сможете видеть историю ваших транзакций.
Email address provided is not a valid email!	Предоставленный адрес электронной почты не является допустимым электронным письмом!	Указан недействительный адрес электронной почты!

Что дает гибридная технология перевода?

- быструю автоматическую настройку на основе translation memories заказчика,
- терминологическую точность перевода, а также единство стиля,
- получение дополнительных полезных данных – двуязычного терминологического словаря

Разница в качестве между гибридным машинным переводом и переводом, выполненным человеком, составляет 25-40 % (== объем постредактирования)

Объемы данных

Минимальный объем данных для тренировки гибридной технологии:

- 50 000 сегментов из ТМ,
- не менее 500 000 слов в исходном языке.

Спасибо!

www.promt.ru