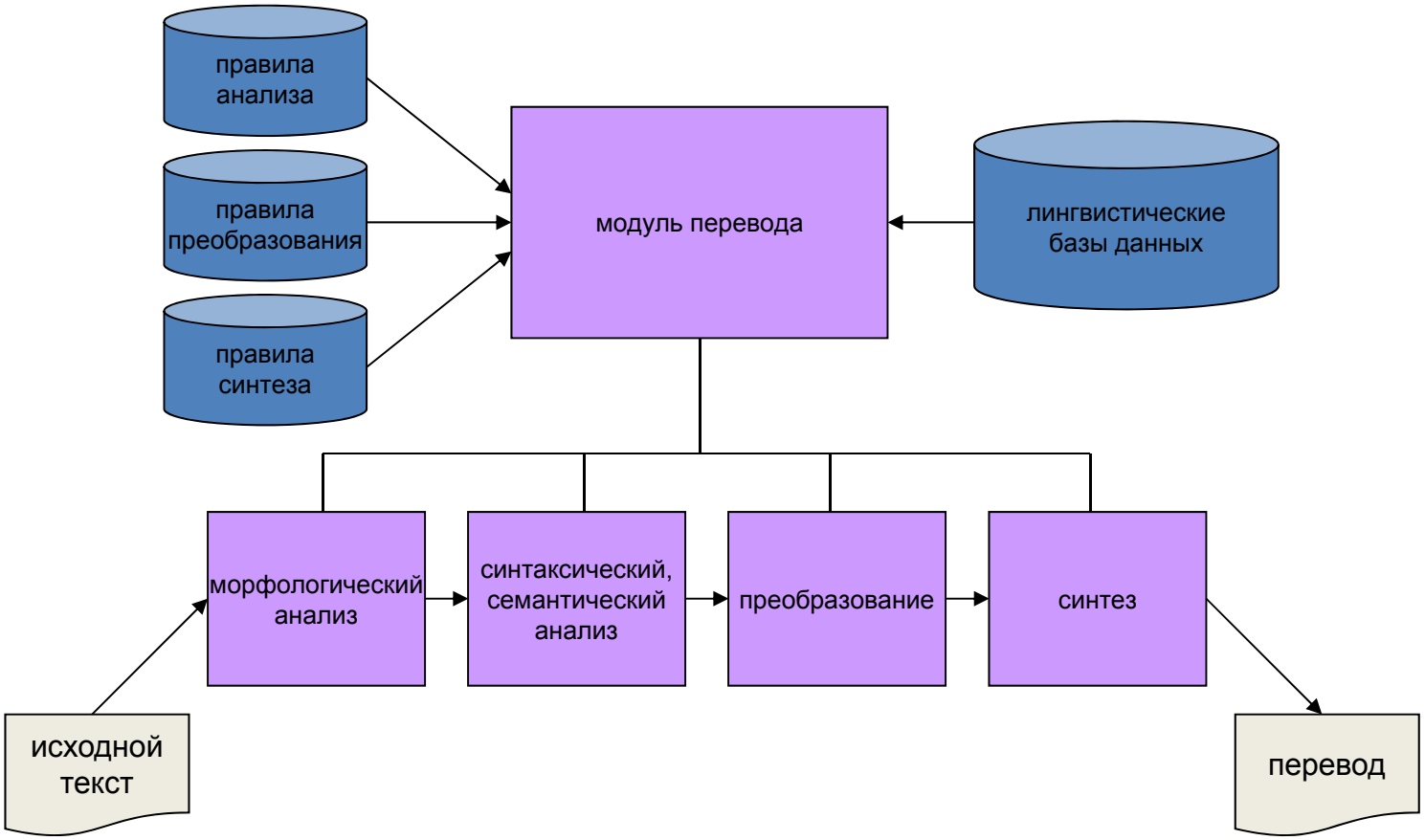
Машинный перевод по правилам и без, или Зачем нужна гибридная технология перевода

Типы систем машинного перевода

- Системы, основанные на правилах (rule-based)
- Статистические системы (системы, основанные на параллельных двуязычных корпусах)

Rule-based системы



Машинный перевод по правилам и без

Компоненты rule-based систем

Лингвистические базы данных

- двуязычные словари
- морфологические таблицы
- списки префиксов
- базы имен

Особенности rule-based систем

Преимущества

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

Недостатки

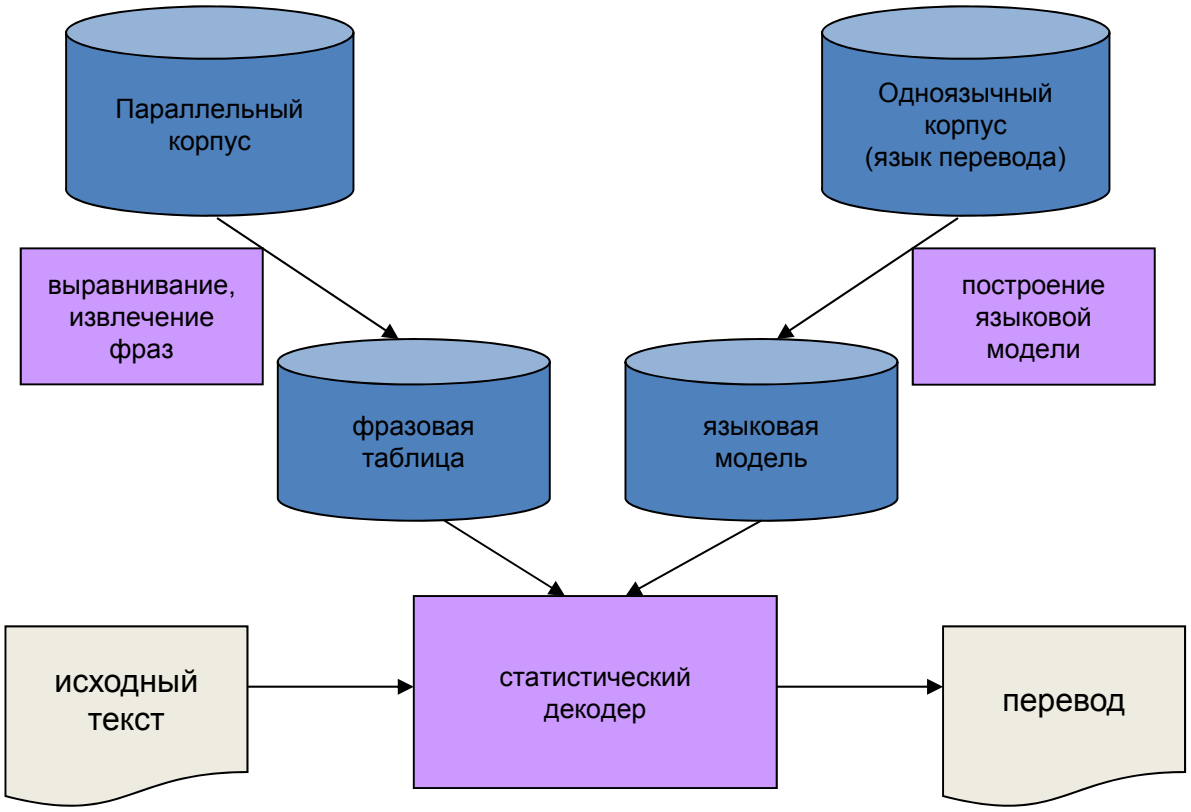
- трудоемкость и длительность разработки
- необходимость поддерживать и актуализировать лингвистические базы данных
- «машинный акцент» при переводе

Rule-based перевод

Since the Desert One debacle, the United States has poured vast resources into its special forces.

Начиная с разгрома Пустыни Один, Соединенные Штаты вылили обширные ресурсы в свой спецназ.

Статистические системы



Исходный текст – это «зашифрованный» перевод, который нужно декодировать

Компоненты статистических систем

Фразовая таблица – таблица соответствий фраз исходного корпуса и корпуса переводов с некоторыми статистическими коэффициентами.

фрагмент фразовой таблицы

исходная фраза	*перевод*	*статистические коэффициенты*
can download it at	можете загрузить его по адресу	1 0.032 0.5 0.01
company	компания	0.39 0.19 0.12 0.11
company	компания-эмитент	0.85 0.42 0.01 0.05
company	кредитной	0.01 0.01 0.01 0.01
company	название компании	0.11 0.13 0.01 0.01
company	организации , выпустившей его	1 0.05 0.01 0.37
compare all of	сравнение всех	1 0.04 0.5 0.03
compare all of	сравните все	0.33 0.01 0.5 0.07

Компоненты статистических систем

Языковая модель – набор n-грамм (последовательностей словоформ длины n) из корпуса текстов.

фрагмент языковой модели

статистический коэффициент	*n-грамма*
-4.697978	в ознаменование
-4.697978	в оказание
-0.766904	<s> метод отправки
-0.508603	новые календари </s>
-0.528649	в календарь </s>
-0.988104	кворума старейшин президентом миссии
-1.048399	а также президентом мексиканской

Особенности статистических систем

Преимущества

- Быстрая настройка
- Легко добавлять новые направления перевода
- Гладкость перевода

Недостатки

- «Дефицит» параллельных корпусов
- Многочисленные грамматические ошибки
- Нестабильность перевода

Статистический перевод

Medvedev is to blame

Медведев виноват

Obama is to blame

Обама не виноват

Гибридные технологии перевода

Задача

Совместить достоинства двух основных подходов и нивелировать их недостатки.

Подходы к созданию гибридных систем

- Интеграция лингвистических правил в статистические системы
- Интеграция статистических методов в rule-based системы

Интеграция лингвистических правил в статистические системы

- Синтаксическая и морфологическая разметка корпусов для обучения
- Применение правил для идентификации и перевода именованных сущностей
- Разбиение сложных слов
- Применение правил для идентификации и перевода отдельных синтаксических конструкций

Решение частных задач не исправляет фундаментальные недостатки статистического перевода

Интеграция статистических методов в rule-based системы

Проблема

Rule-based системы имеют недостатки алгоритмов анализа и синтеза, которые постоянно воспроизводятся при переводе.

Решение

Научить систему автоматически корректировать эти недостатки.

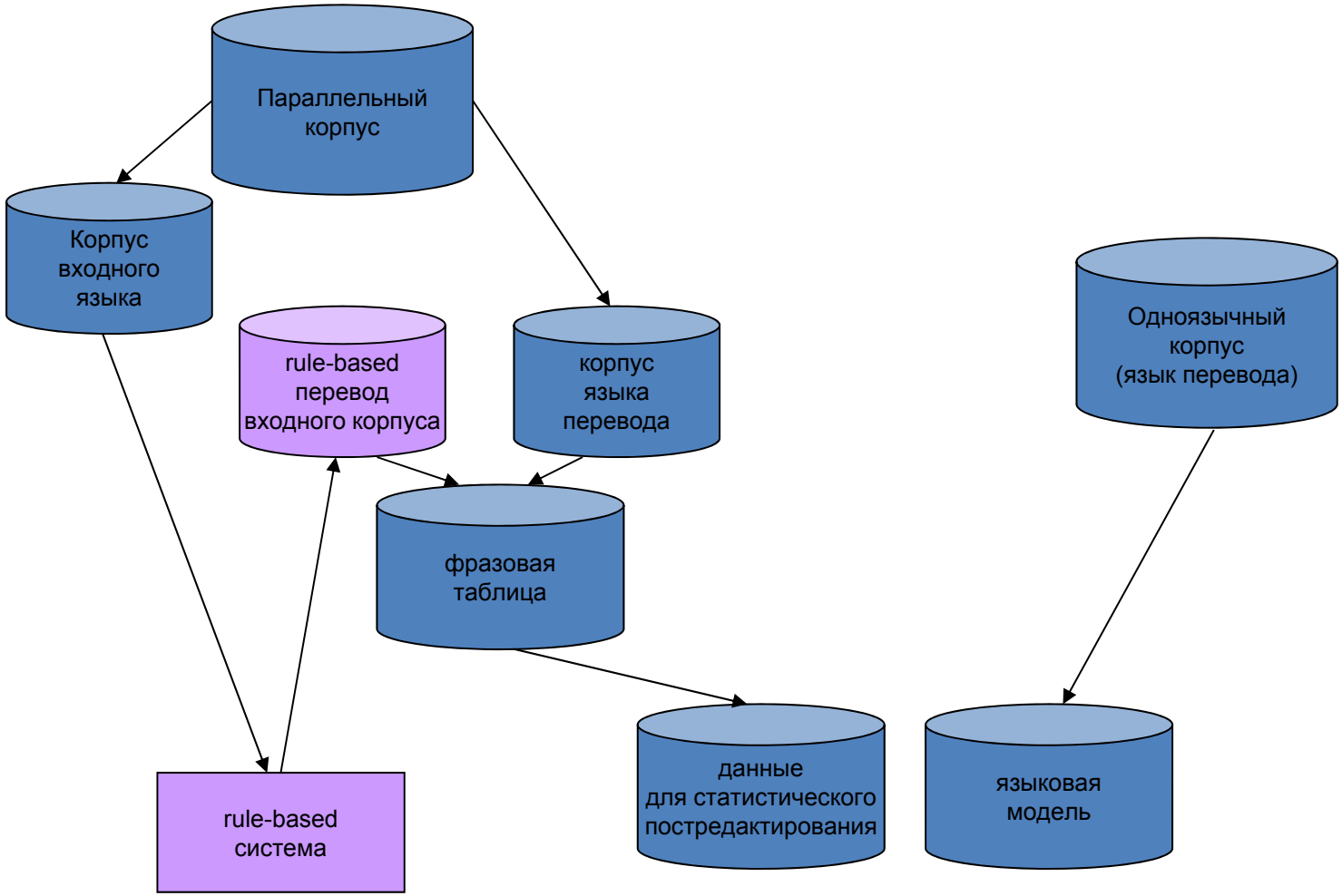
- Система статистического постредактирования

Гибридная технология перевода PROMT

- Этап обучения системы с помощью статистических методов

- Использование полученных на этапе обучения данных в процессе перевода

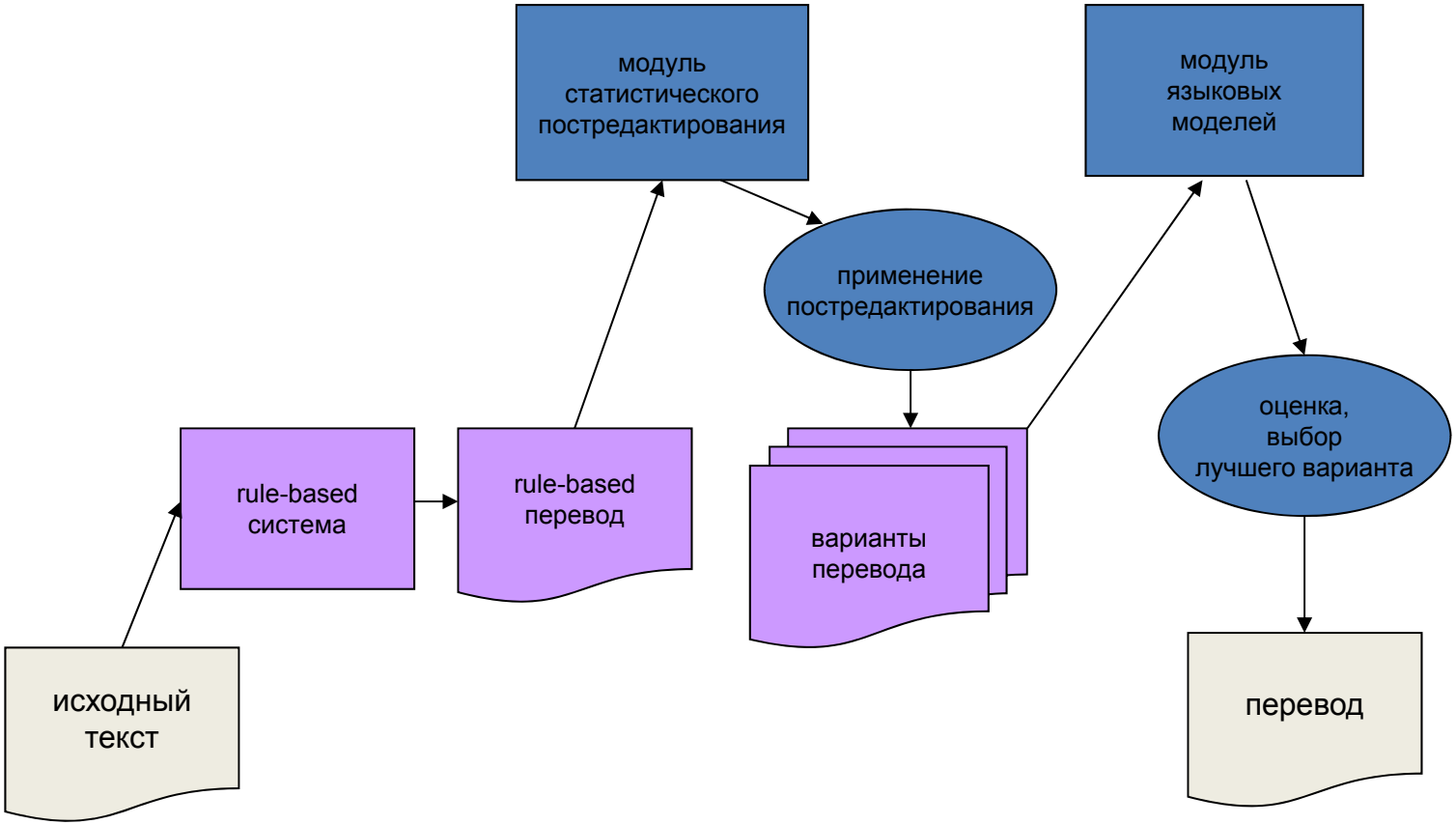
Обучение системы



Зачем нужна гибридная технология перевода

PROMT®

Процесс перевода



Гибридный перевод

- Четкая синтаксическая структура перевода
- Стабильность
- Гладкость

Гибридный перевод

Исходный текст

Before proceeding further, every effort was made by senior staff to ensure that a friendly atmosphere prevailed.

Rule-based перевод

Прежде, чем продолжиться далее, каждое усилие было приложено руководящим персоналом, чтобы гарантировать, что преобладала дружественная атмосфера.

Гибридный перевод

Прежде чем продолжить, руководящий персонал предпринял все усилия, чтобы преобладала дружественная атмосфера.

Гибридный перевод

Исходный текст

In the dialog box that opens, specify the necessary export settings.

Rule-based перевод

В диалоговом окне, которое открывается, определите необходимые настройки экспорта.

Гибридный перевод

В открывшемся диалоговом окне укажите требуемые параметры экспорта.

Оценка машинного перевода

Большинство существующих метрик автоматической оценки МП основаны на сравнении с человеческим эталоном.

Необходима метрика оценки МП в отсутствие эталона.

- автоматический перевод контента сетевых ресурсов
- оценка изменений в технологии перевода

Метрики оценки машинного перевода

BLEU Score
Meteor
TER (Translation Error Rate)

Совпадение n-грамм в машинном переводе и эталоне

машинный перевод	**эталон**
Нажмите «Да» в окне сообщения, которое появляется.	Нажмите «Да» в открывшемся окне сообщения.

совпавшие n-граммы	
1	Нажмите ; «Да» ; в ; окне
2	Нажмите «Да» ; «Да» в
3	Нажмите "Да" в

Оценка машинного перевода без человеческого эталона

Статистическая языковая модель (Perplexity)

Perplexity – величина, обратно пропорциональная вероятности.

Исходный текст
Click Yes in the message window that appears.

Машинный перевод 1
Нажмите «Да» в окне сообщения, которое появляется.
(Perplexity = 842)

Машинный перевод 2
Нажмите «Да» в открывшемся окне сообщения.
(Perplexity = 438)

Оценка машинного перевода без человеческого эталона

Perplexity не учитывает особенности входного текста и грамматическую структуру перевода.

Исходный текст
Click Yes in the message window that appears.

Машинный перевод 1
Нажмите «Да» в открывшемся окне сообщения.
(Perplexity = 438)

Машинный перевод 2
Нажмите «Да»
(Perplexity = 145)

Оценка машинного перевода без человеческого эталона

Необходимо разработать комплексную метрику оценки на основе характеристик входного и выходного текста:

- длина текста (количество слов)
- вероятности перевода слов и фраз
- совпадение чисел, дат и т.п.
- оценка с помощью языковой модели
- морфологические, синтаксические признаки

Спасибо!

Александр Молчанов

Компания ПРОМТ

Alexander.Molchanov@promt.ru