Основы обработки текстов

Лекция 10

Машинный перевод

План

- Применение машинного перевода
- Сложности перевода
  - Типология
  - Различия языков
- Классический подход
- Статистический подход
  - Модель зашумленного канала
  - Выравнивание
  - Тренировка моделей
  - Декодирование
  - Методы оценки

Применение машинного перевода

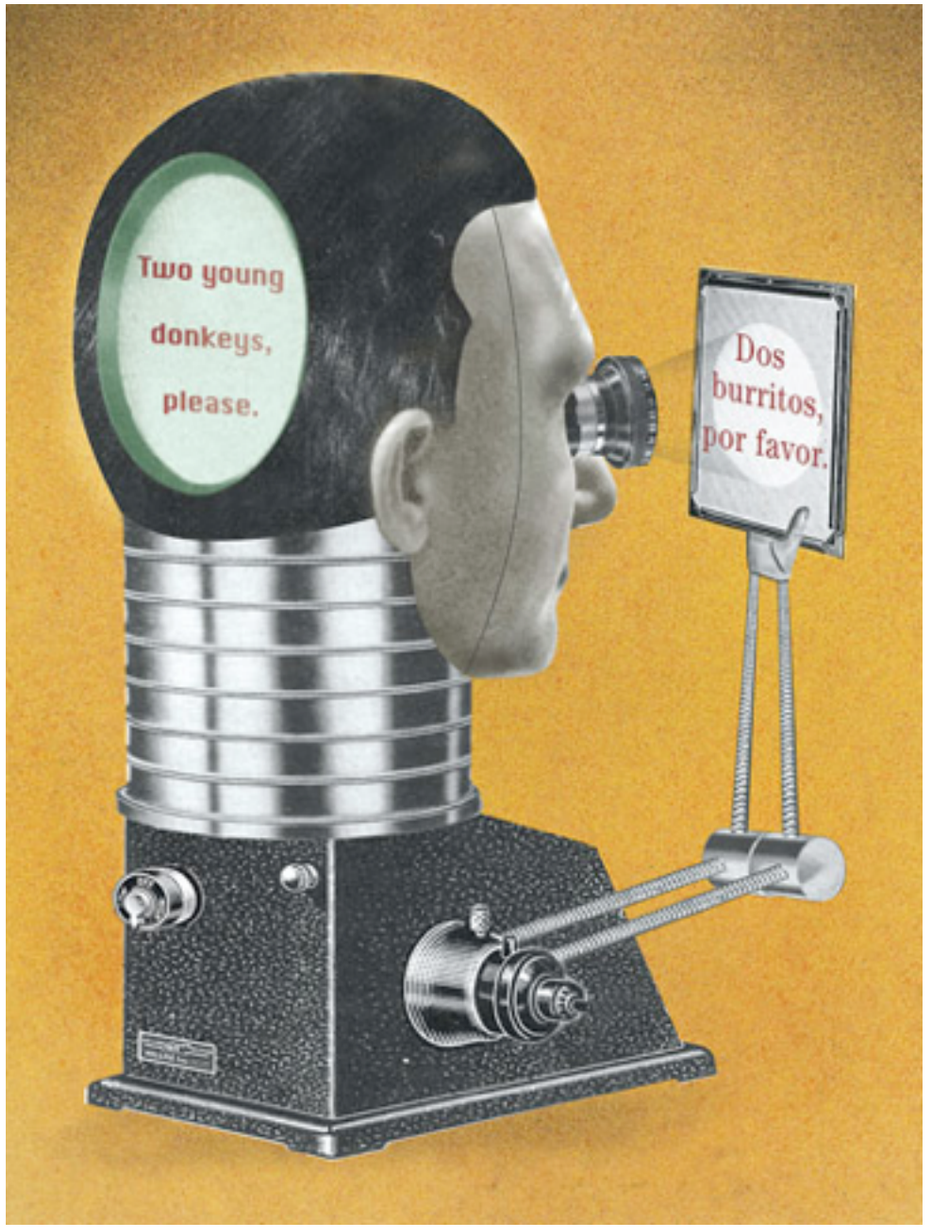
- Задачи, где достаточно грубого перевода
  - Задачи извлечения информации
  - Перевод Веб-страниц
  - e-mail
- Задачи, где результат перевода может быть исправлен
  - Помощь переводчику
- Перевод подмножеств языка
  - FAHQT (Fully Automatic High Quality Translation)

Где машинный перевод недостаточно хорош

- Художественная литература
- Разговорный язык
- Медицинский перевод в больницах
- Звонки в службу спасения

Сложность перевода

- Некоторые аспекты языков схожи, некоторые различны
- Различия в
  - –морфологии
  - –лексике
  - –структуре



Морфология

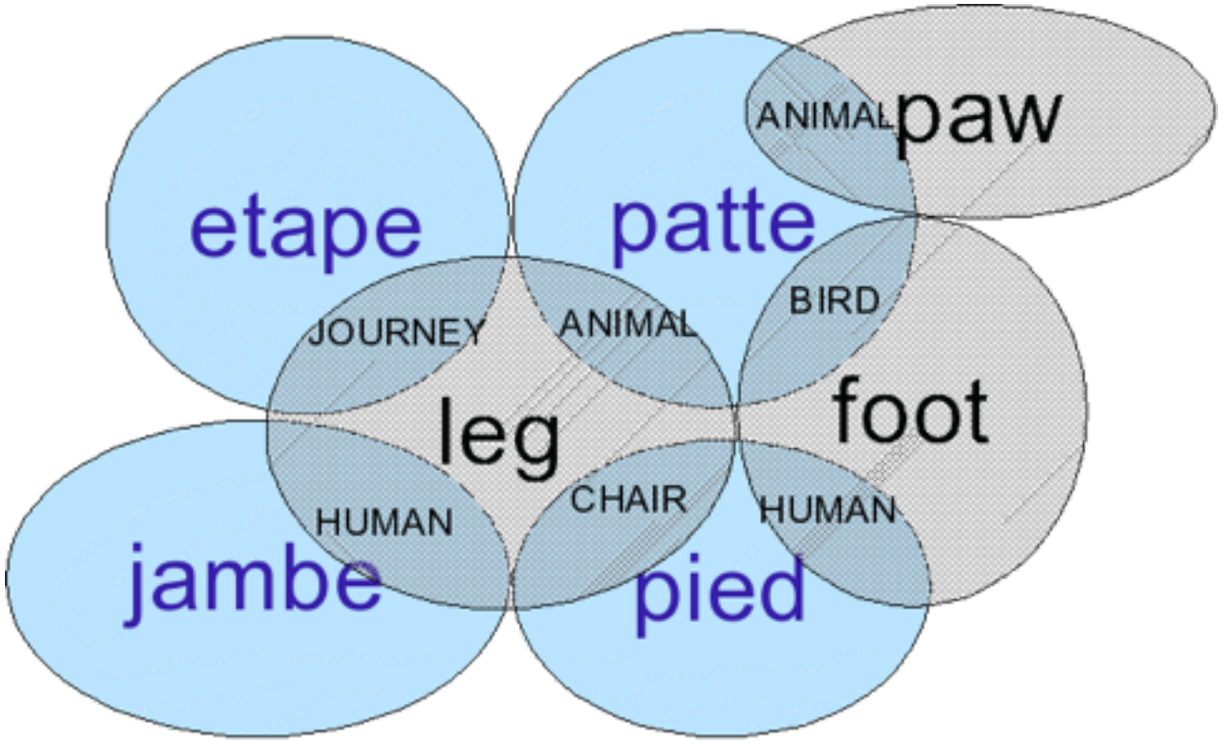
- Морфема
  - минимальная значимая единица языка
  - слово = морфема + морфема + морфема + ...
- Аффиксы
  - Префикс: undo
  - Суффикс: looking
  - Инфикс: hingi (занимать) - humingi (заемщик) (Тагальский язык)
  - Циркумфикс: sagen (сказать) - gesagt (сказал) (Немецкий)

Морфологические различия

- Изолирующие языки
  - Каждое слово состоит из одной морфемы (Вьетнамский)
- Полисинтетические языки
  - слово состоит из множества морфем (Чукотский: Тымэйӈылевтпыгтыркын - У меня сильно болит голова)
- Аглютинативные
  - Морфемы несут определенные значения (Турецкий)
- Флективные
  - Морфемы имеют несколько значений (Русский: “хороший” - им. падеж, ед. число, муж. род)

Лексические различия

- Семантические особенности:
  - В корейском нет слов брат/сестра, есть старший/младший брат/сестра
  - В чукотском около 20 слов для снега
- Английский vs французский



Синтаксические различия

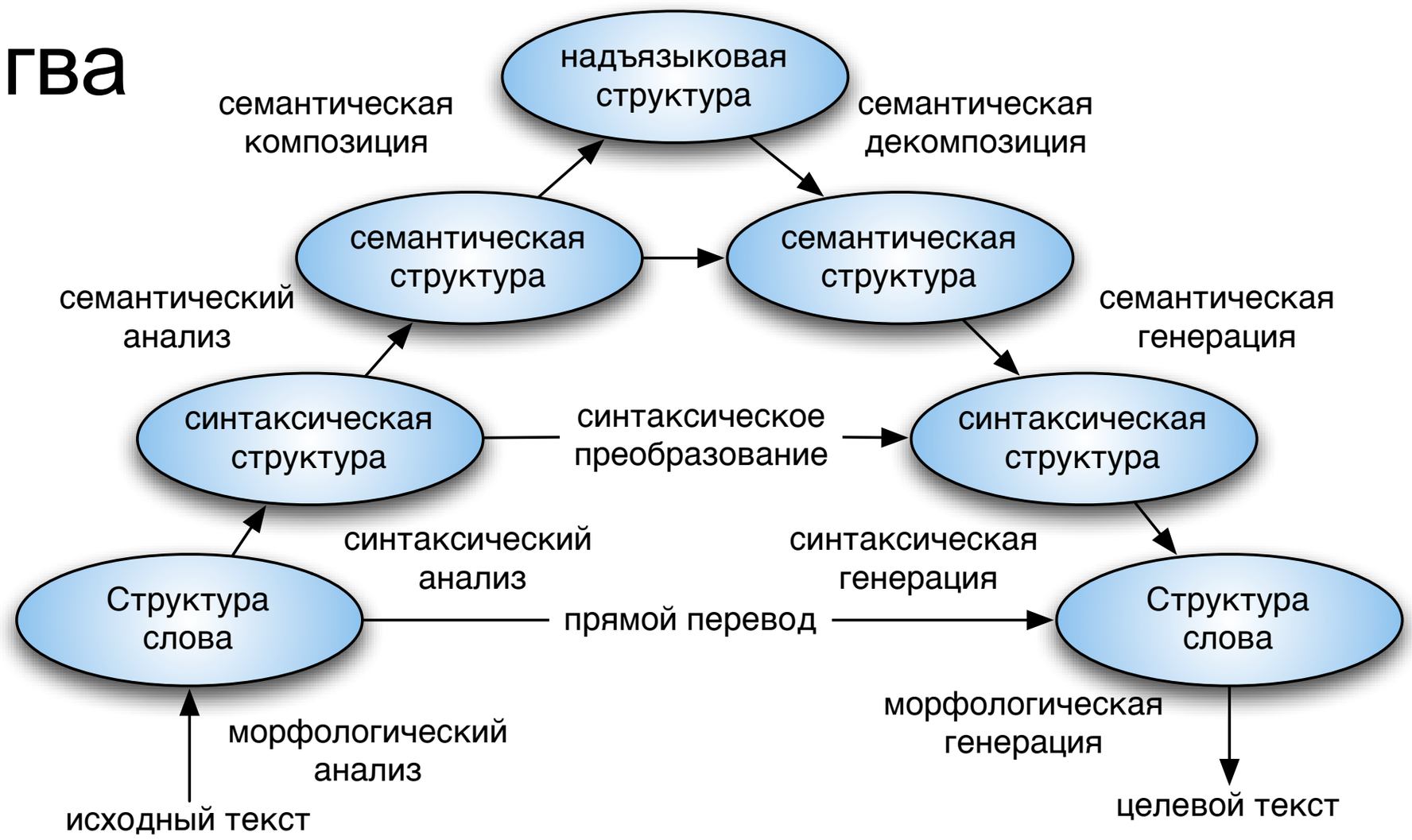
- СГО (Субъект-Глагол-Объект)
  - Английский, Немецкий
  - I am in Moscow
- СОГ
  - Японский, Корейский
  - 저는 모스크바에 있습니다 (Я в Москве нахожусь)
- ГСО
  - Ирландский, классический Арабский

Границы

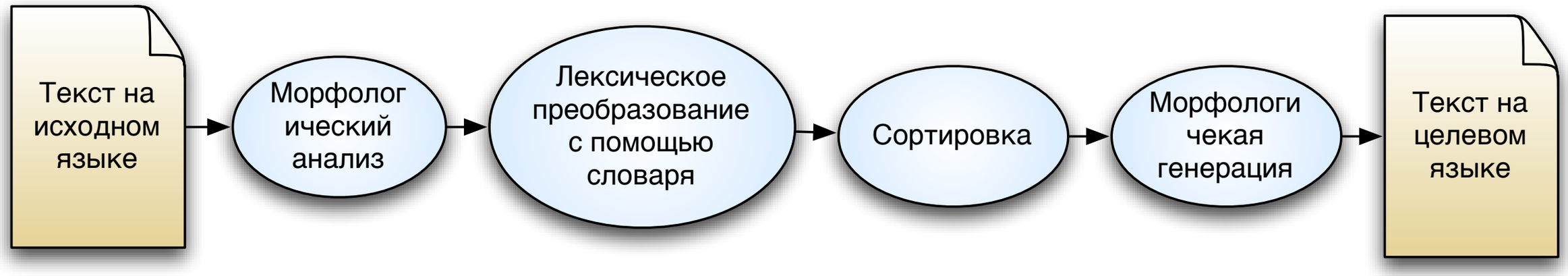
- Языки в которых не выделены границы слов:
  - Китайский, Японский, Тайский, Въетнамский
- Предложения в некоторых языках больше похожи на параграфы
  - Китайский, современный Арабский

Классические подходы

- Прямой перевод
- Преобразование
- Интерлингва



Подход 1: Прямой перевод



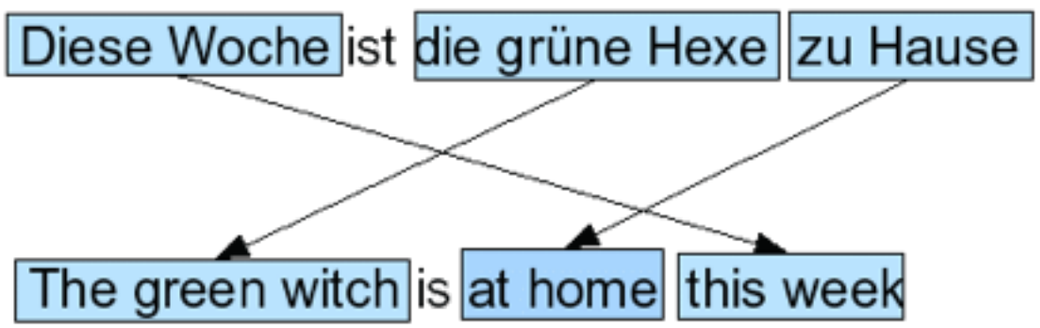
- Последовательный перевод каждого слова
- Не используется никакие структуры кроме морфологии
- После перевода слов, делается сортировка

Пример

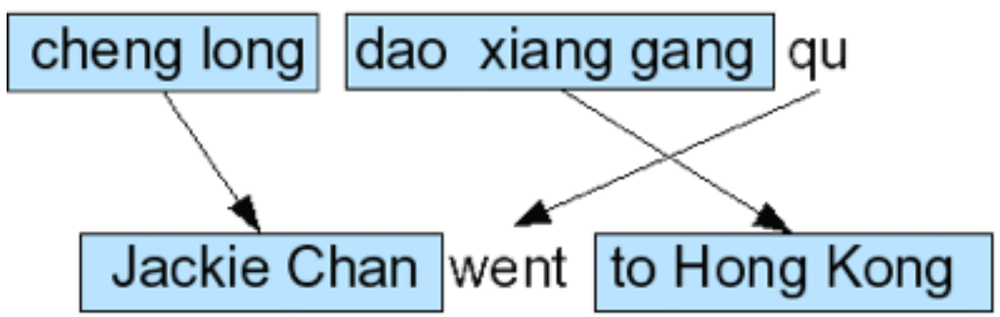
Input:	Mary didn't slap the green witch
After 1: Morphology	Mary DO-PAST not slap the green witch
After 2: Lexical Transfer	Maria PAST no dar una bofetada a la verde bruja
After 3: Local reordering	Maria no dar PAST una bofetada a la bruja verde
After 4: Morphology	Maria no dió una bofetada a la bruja verde

Проблемы

- Сложные перестановки
  - –термины
  - –длинные дистанции
- Немецкий



- Китайский

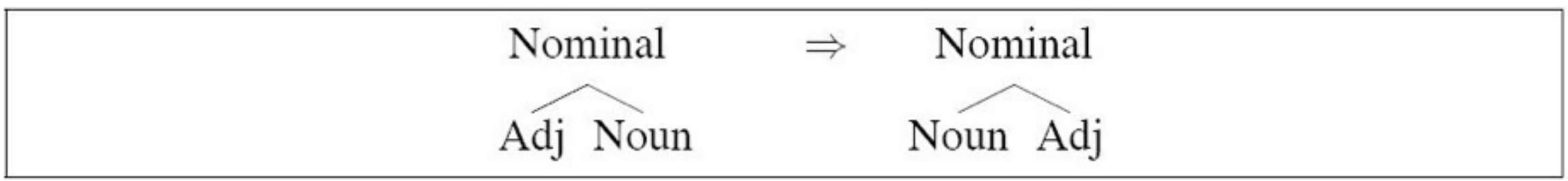


Подход 2: Преобразование

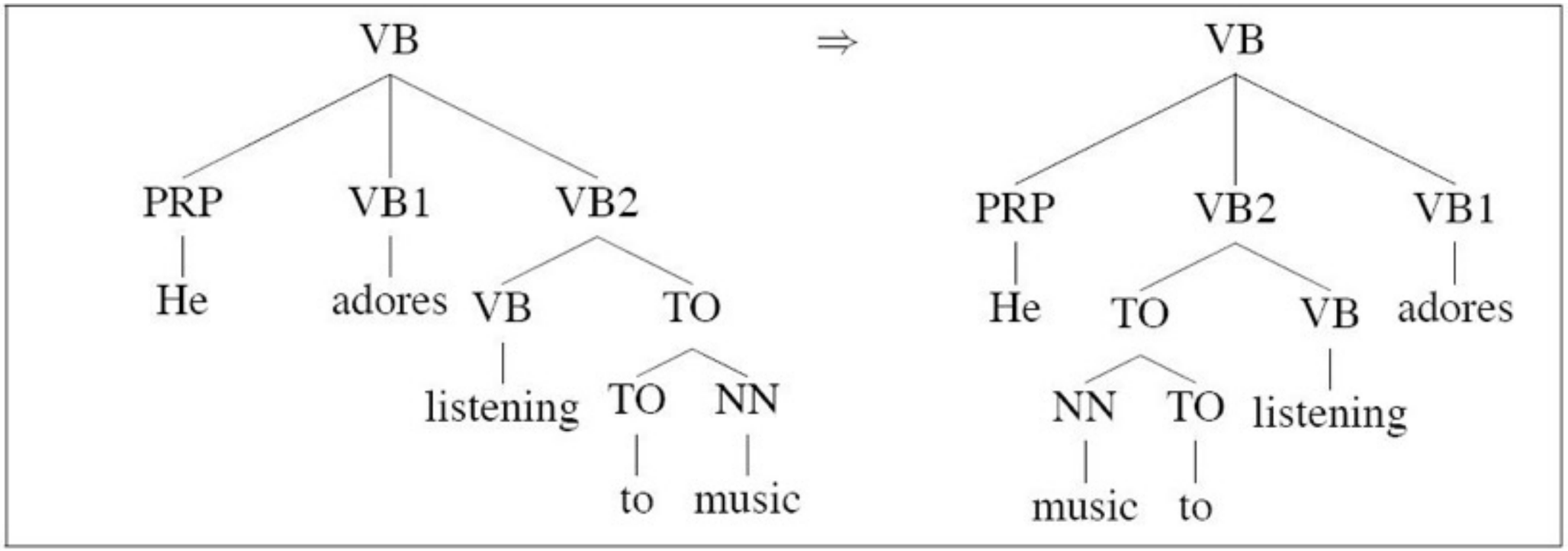
- Применение знаний о различиях в языках
- Шаги
  - Анализ: синтаксический разбор исходного предложения
  - Преобразование: правила преобразования разбора в разбор на целевом языке
  - Генерация предложения на целевом языке

Пример

- Английский: прилагательное существительное
- Французский: существительное прилагательное
- Не всегда
- Правило



Правила преобразования



	English to Spanish:		
1.	NP → Adjective$_1$ Noun$_2$	⇒	NP → Noun$_2$ Adjective$_1$
	Chinese to English:		
2.	VP → PP[+Goal] V	⇒	VP → V PP[+Goal]
	English to Japanese:		
3.	VP → V NP	⇒	VP → NP V
4.	PP → P NP	⇒	PP → NP P
5.	NP → NP$_1$ Rel. Clause$_2$	⇒	NP → Rel. Clause$_2$ NP$_1$

Systran: комбинирование подходов

- Анализ
  - Морфологический, определение частей речи
  - Группировка (chunking)
  - Разбор некоторых зависимостей
- Преобразование
  - перевод идиом
  - Разрешение лексической многозначности
  - назначение предлогов на основе моделей управления глаголов
- Синтез
  - Применения большого двуязычного словаря
  - сортировка
  - морфологическая генерация

Проблемы

- Грамматика и лексика содержат много специфики
- Трудно сделать и еще труднее поддерживать

Интерлингва

- Пример системы: ABBYY Compreno
- Идея: Вместо использования правил преобразования между языками использовать значение предложения
- Шаги
  - Перевести исходное предложение в представление его значения
  - Сгенерировать целевое предложение из значения

Интерлингва

Mary did not slap the green witch

$$\begin{bmatrix} \text{EVENT} & \text{SLAPPING} \\ \text{AGENT} & \text{MARY} \\ \text{TENSE} & \text{PAST} \\ \text{POLARITY} & \text{NEGATIVE} \\ \text{THEME} & \begin{bmatrix} \text{WITCH} & \\ \text{DEFINITENESS} & \text{DEF} \\ \text{ATTRIBUTES} & \begin{bmatrix} \text{HAS-COLOR} & \text{GREEN} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Проблемы

- Разные понятия в языках
  - 20 типов снега в Чукотском
  - Не нужны для англо-русского перевода
- Всесторонний анализ семантики и представление знаний
  - Возможно сделать только для специфичных подмножеств языка

Статистический машинный перевод

- Идеи:
  - Использование параллельных текстов
  - Перевод по фразам
  - Сортировка результата

Розеттский камень:
- древнегреческий
- древнеегипетский
- древнеегипетские иероглифы

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]
Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . ??? 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat . ???
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

методом исключения

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]

Перевести: **farok crrrok hihok yorok clok kantok ok-yurp**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Родственное слово?

Перевод на основе параллельных корпусов

Перевод с Центаврианского на Арктуранский [Knight, 1997]
Задание: упорядочить: {jjat, arrat, mat, bat, oloat, at-yurp}

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

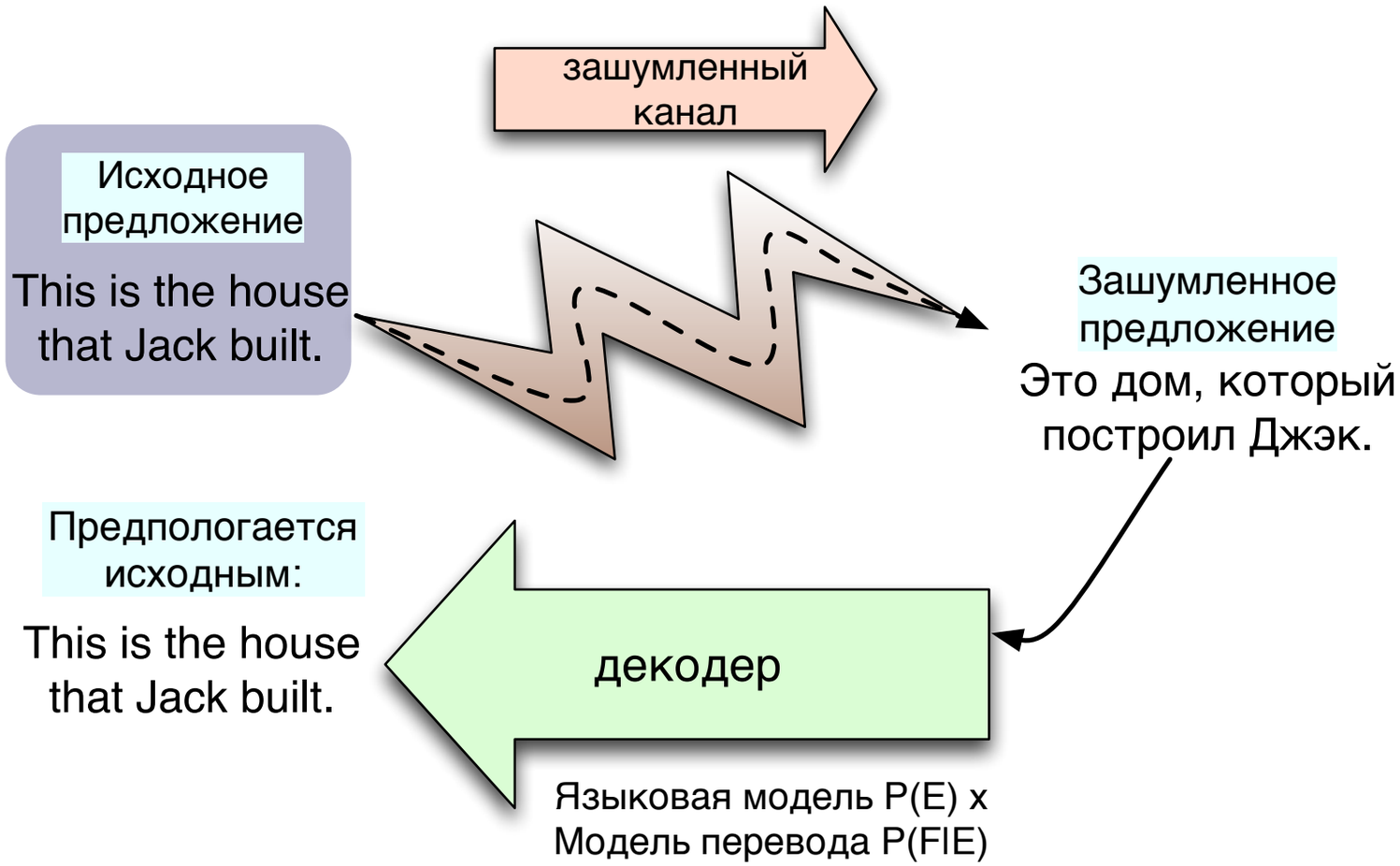
Перевод на основе параллельных корпусов

В действительности это англо-испанский перевод

Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Модель зашумленного канала



- Байесовская модель

$$\hat{E} = \underset{E \in English}{\arg\max}\ P(F|E)P(E)$$

Модель перевода

Языковая модель

Машинный перевод

- Языковая модель
  - N-граммы
  - СКС грамматики
- Модель перевода
- Декодер

Модель перевода на основе фраз

- P(F|E)
- Разбиваем E на фразы
- Переводим каждую фразу из E во фразу на другом языке, запоминая **вероятность перевода**
- Находим наиболее вероятную последовательность фраз F

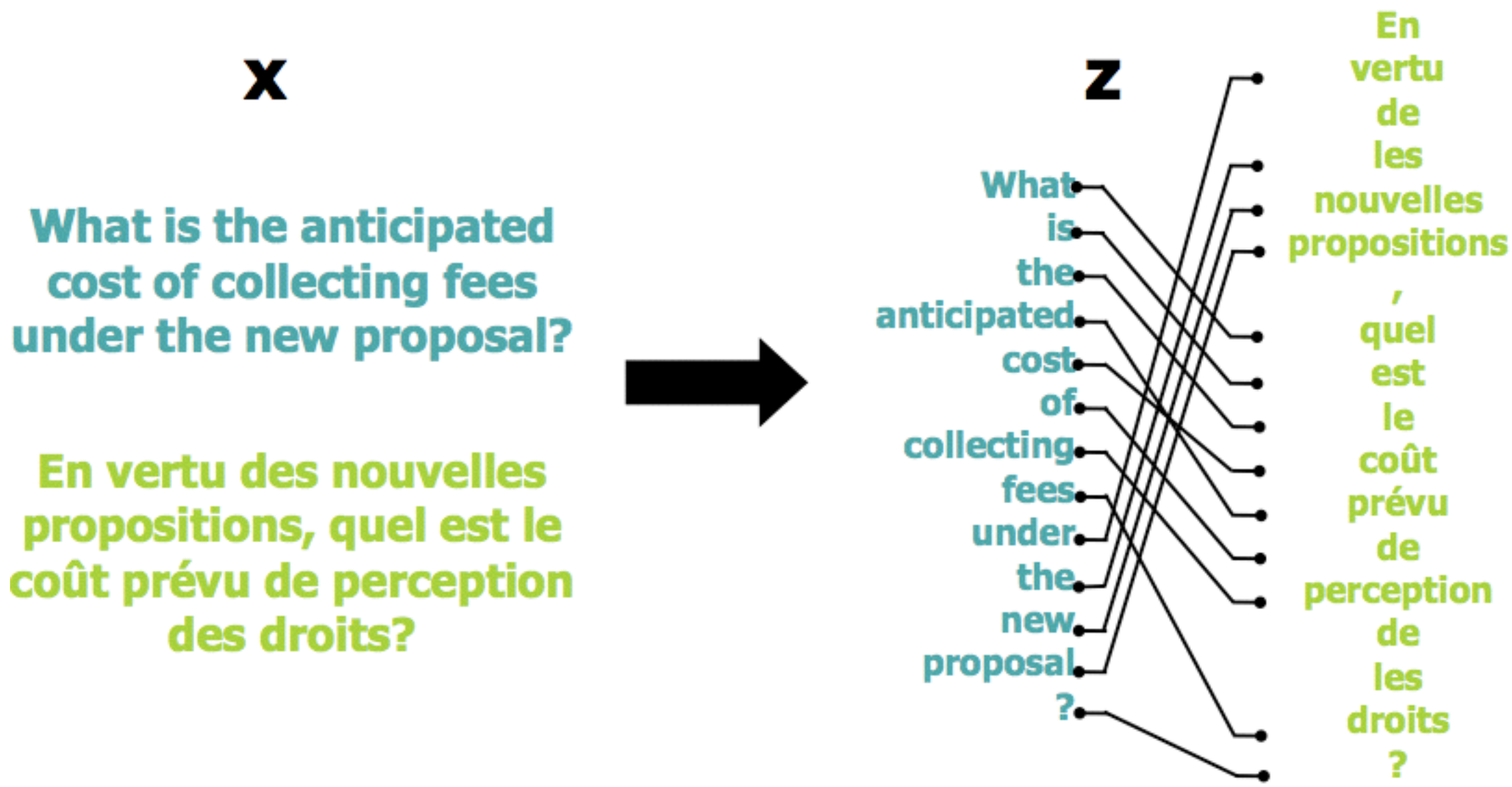
Вероятность перевода

- Пусть есть параллельные тексты с указанием соответствия между фразами в E и F (см. далее).
- Тогда вероятность перевода можно оценить на основе метода максимального правдоподобия

$$\phi(\overline{f}, \overline{e}) = \frac{count(\overline{f}, \overline{e})}{\sum_{\overline{f}}(\overline{f}, \overline{e})}$$

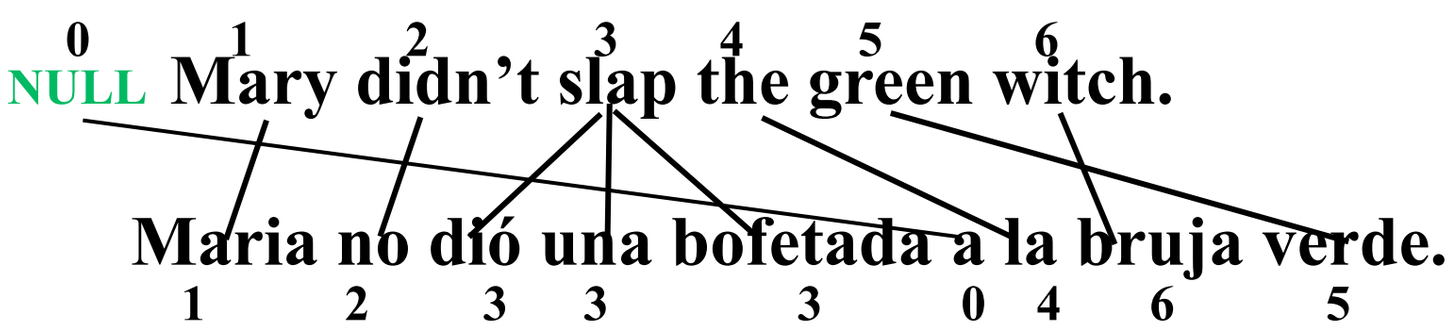
Выравнивание слов

- Сначала выравнивают слова
- Вход: пары предложение-перевод



Выравнивание один ко многим

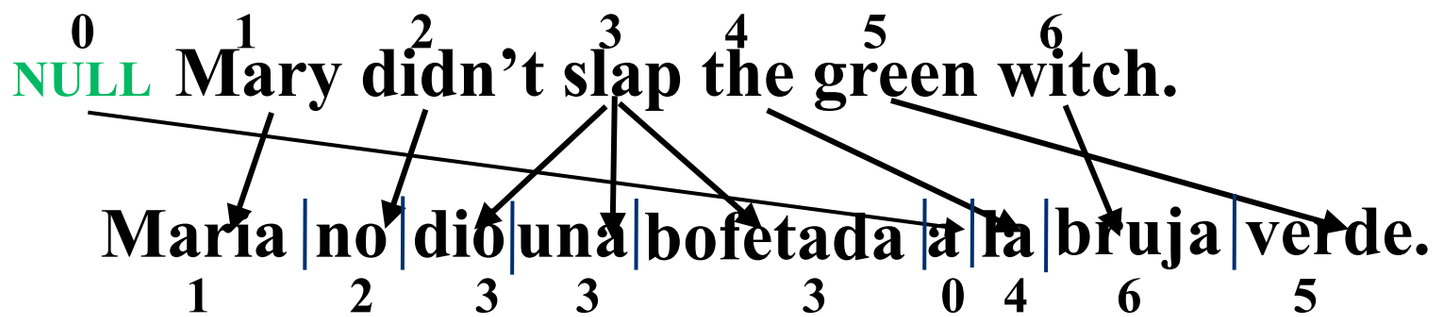
- Для простоты предположим что
  - слово из F соответствует одному слову в E
  - но слово из E может соответствовать нескольким словам в F
- Некоторые слова в F могут соответствовать элементу NULL в E
- Тогда выравнивание можно задавать вектором



Модель IBM Model 1

- Первая самая простая модель предложенная в основополагающей статье [Brown et. al. 1993]
- Предлагает простую генерирующую модель для получения F из E=$e_1, e_2, \ldots e_I$
  - Выбрать длину J предложения F=$f_1, f_2, \ldots f_J$
  - Выбрать выравнивание $A=a_1, a_2, \ldots a_J$
  - Получить F из E

Пример


0 1 2 3 4 5 6
NULL Mary didn't slap the green witch.
María no dió una bofetada a la bruja verde.
1 2 3 3 3 0 4 6 5

Подсчет P(F|E)

- Обозначения
  - Длина предложения на языке E равна I
  - Длина предложения на языке F равна J
  - Вероятность длины предложения в F равна P(J|E)
- Model 1: Предположим что все выравнивания A равновероятны (их $(I+1)^J$)
- Тогда условная вероятность (выравниваний):

$$P(A \mid E) = P(A \mid E, J) P(J \mid E) = \frac{P(J \mid E)}{(I+1)^J}$$

Подсчет P(F|E)

- Пусть $t(f_x, e_y)$ вероятность перевода слова e_y в слово f_x
- Определим **P(F|E)**

$$P(F \mid E) = \sum_A P(F \mid E, A) P(A \mid E) = \sum_A \frac{P(J \mid E)}{(I+1)^J} \prod_{j=1}^{J} t(f_j, e_{a_j})$$

Декодирование

- Цель: найти наиболее вероятное выравнивание

$$\begin{aligned}\hat{A} &= \underset{\mathrm{A}}{\operatorname{argmax}}\, P(F, A \mid E) \\ &= \underset{A}{\operatorname{argmax}} \frac{P(J \mid E)}{(I+1)^J} \prod_{j=1}^{J} t(f_j, e_{a_j}) \\ &= \underset{A}{\operatorname{argmax}} \prod_{j=1}^{J} t(f_j, e_{a_j})\end{aligned}$$

- Так как различные переводы для каждой позиции j независимы, максимум произведения достигается при максимуме каждого термина

$$a_j = \underset{0 \le i \le I}{\operatorname{argmax}}\, t(f_j, e_i) \quad 1 \le j \le J$$

Тренировка моделей выравнивания

- Если есть выровненный вручную корпус, то можно оценить параметры модели IBM 1 через метод максимального правдоподобия
- Часто такого корпуса нет. В этом случае применяют EM-алгоритм

EM-алгоритм для выравнивания

- Выбираем начальные параметры
- Пока не сойдется выполняем:
  - E-шаг: Вычисляем вероятность всех выравниваний с помощью текущей модели
  - M-шаг: Используем эти вероятности для переоценки значений всех параметров модели

Для сокращения времени работы используется алгоритм динамического программирования

Пример

Тренировочный корпус

green house
casa verde

the house
la casa

Вероятности перевода

	verde	**casa**	**la**
green	1/3	1/3	1/3
house	1/3	1/3	1/3
the	1/3	1/3	1/3

Предполагаем начальные вероятности равными

Вычисляем вероятности выравнивания P(A, F | E)

green house	**green house**	**the house**	**the house**
casa verde	**casa verde**	**la casa**	**la casa**
1/3 X 1/3 = 1/9	**1/3 X 1/3 = 1/9**	**1/3 X 1/3 = 1/9**	**1/3 X 1/3 = 1/9**

Нормализуем P(A | F, E)

$$\frac{1/9}{2/9} = \frac{1}{2} \qquad \frac{1/9}{2/9} = \frac{1}{2} \qquad \frac{1/9}{2/9} = \frac{1}{2} \qquad \frac{1/9}{2/9} = \frac{1}{2}$$

$$P(A|E,F) = \frac{P(A,F|E)}{\sum_A P(A,F|E)}$$

Пример

green house	green house	the house	the house
casa verde	casa verde	la casa	la casa
1/2	1/2	1/2	1/2

Считаем веса переводов

	verde	casa	la
green	1/2	1/2	0
house	1/2	1/2 + 1/2	1/2
the	0	1/2	1/2

Нормализуем и получаем P(f | e)

	verde	casa	la
green	1/2	1/2	0
house	1/4	1/2	1/4
the	0	1/2	1/2

Пример

Вероятности перевода

	verde	casa	la
green	1/2	1/2	0
house	1/4	1/2	1/4
the	0	1/2	1/2

Пересчитываем вероятности выравнивания P(A, F | E)

green house	green house	the house	the house
casa verde	casa verde	la casa	la casa
1/2 X 1/4=1/8	1/2 X 1/2=1/4	1/2 X 1/2=1/4	1/2 X 1/4=1/8

$$P(A, F|E) = \prod_{j=1}^{J} t(f_j|e_{a_j})$$

Нормализуем и получаем P(A | F, E)

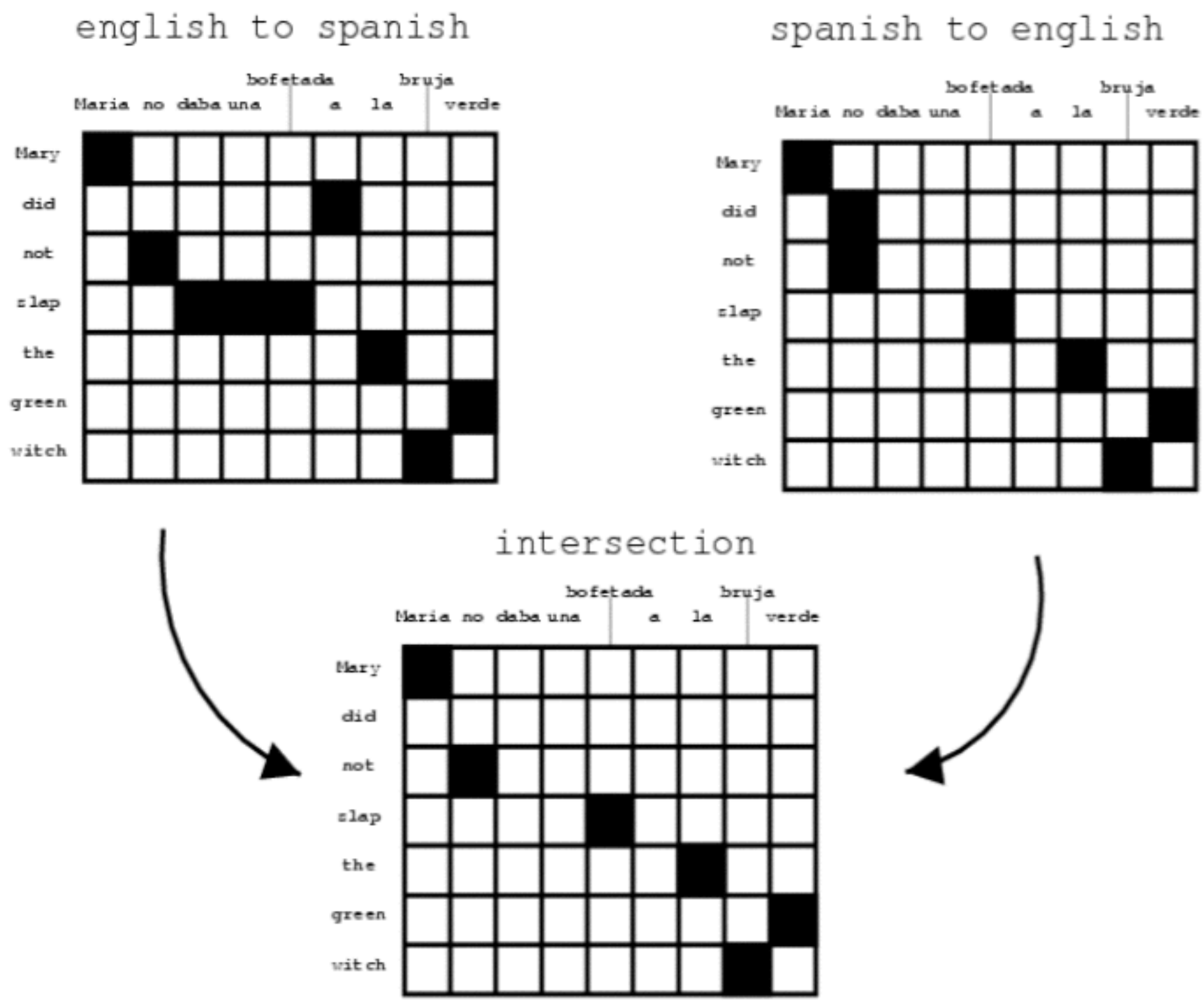
$\frac{1/8}{3/8} = \frac{1}{3}$ $\frac{1/4}{3/8} = \frac{2}{3}$ $\frac{1/4}{3/8} = \frac{2}{3}$ $\frac{1/8}{3/8} = \frac{1}{3}$

Продолжаем алгоритм до сходимости или ограниченное число шагов

Выравнивание фраз

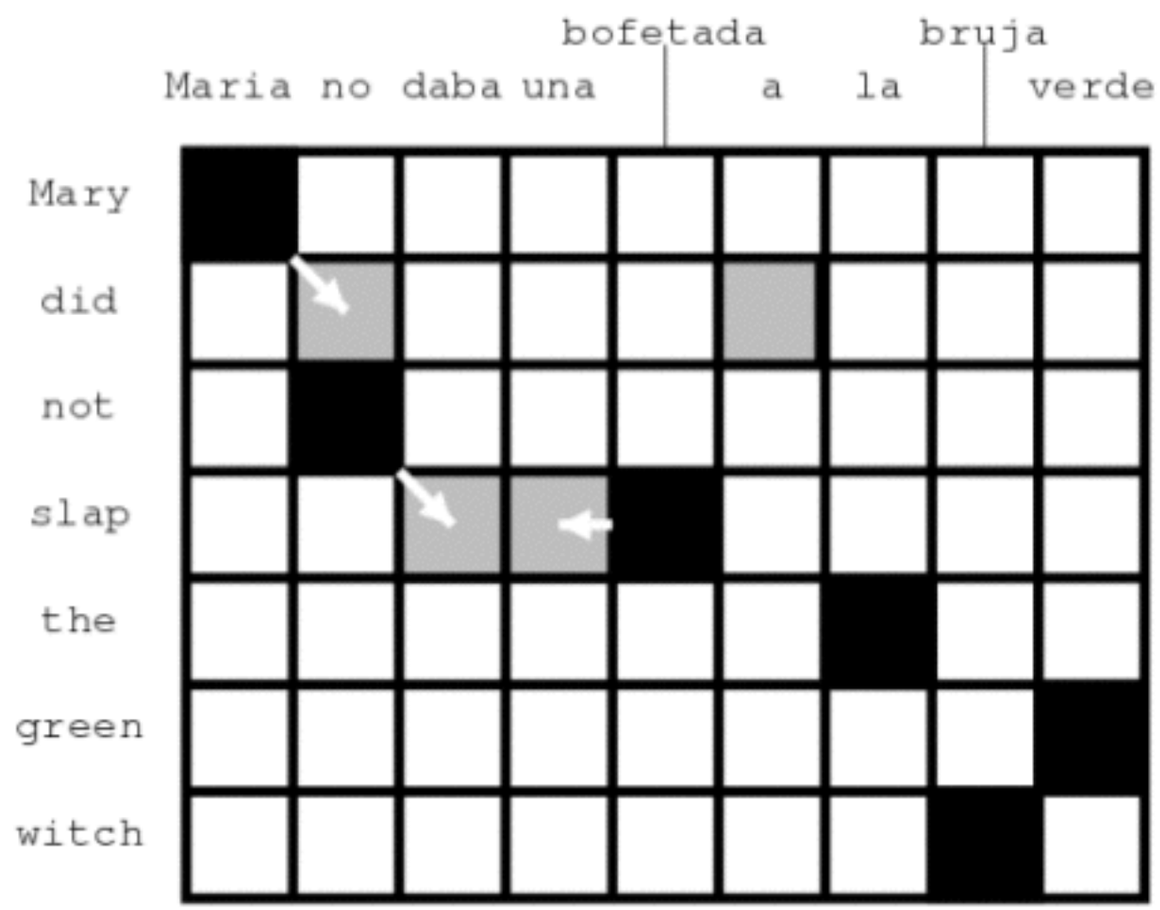
- Мы обсудили как выравнивать слова и переводить текст по словам
- Теперь перейдем к фразам

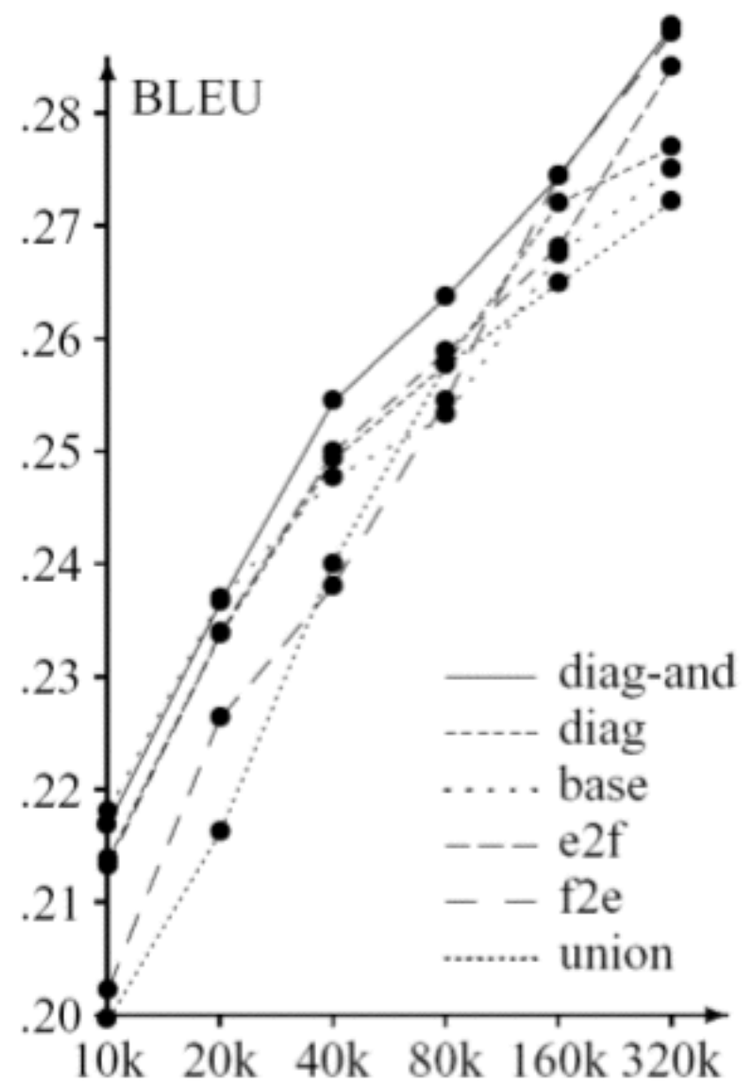
Выравнивание фраз



Выравнивание фраз

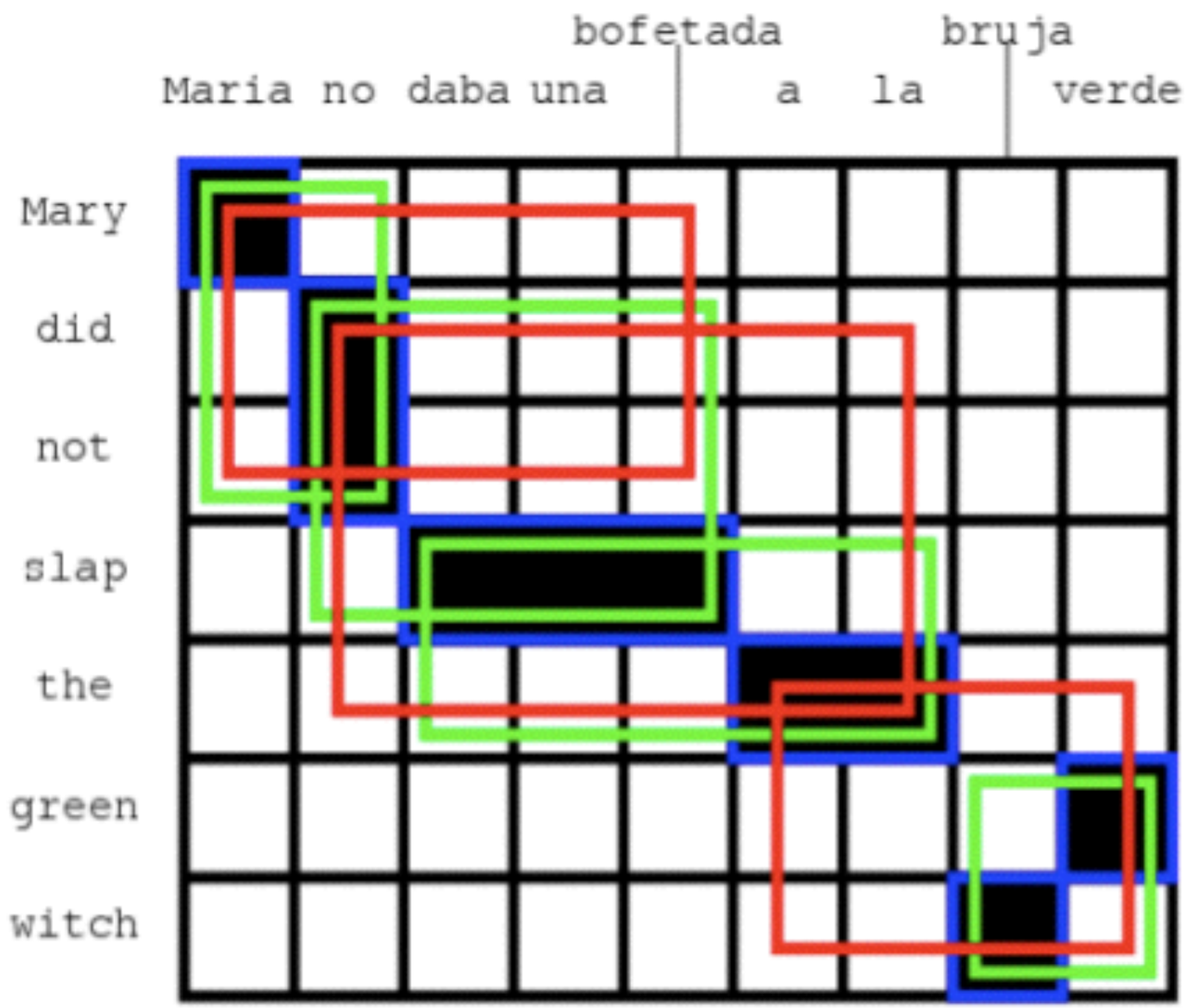
- Существует несколько эвристических методов выравнивания фраз по матрице пересечений





Извлечение фраз

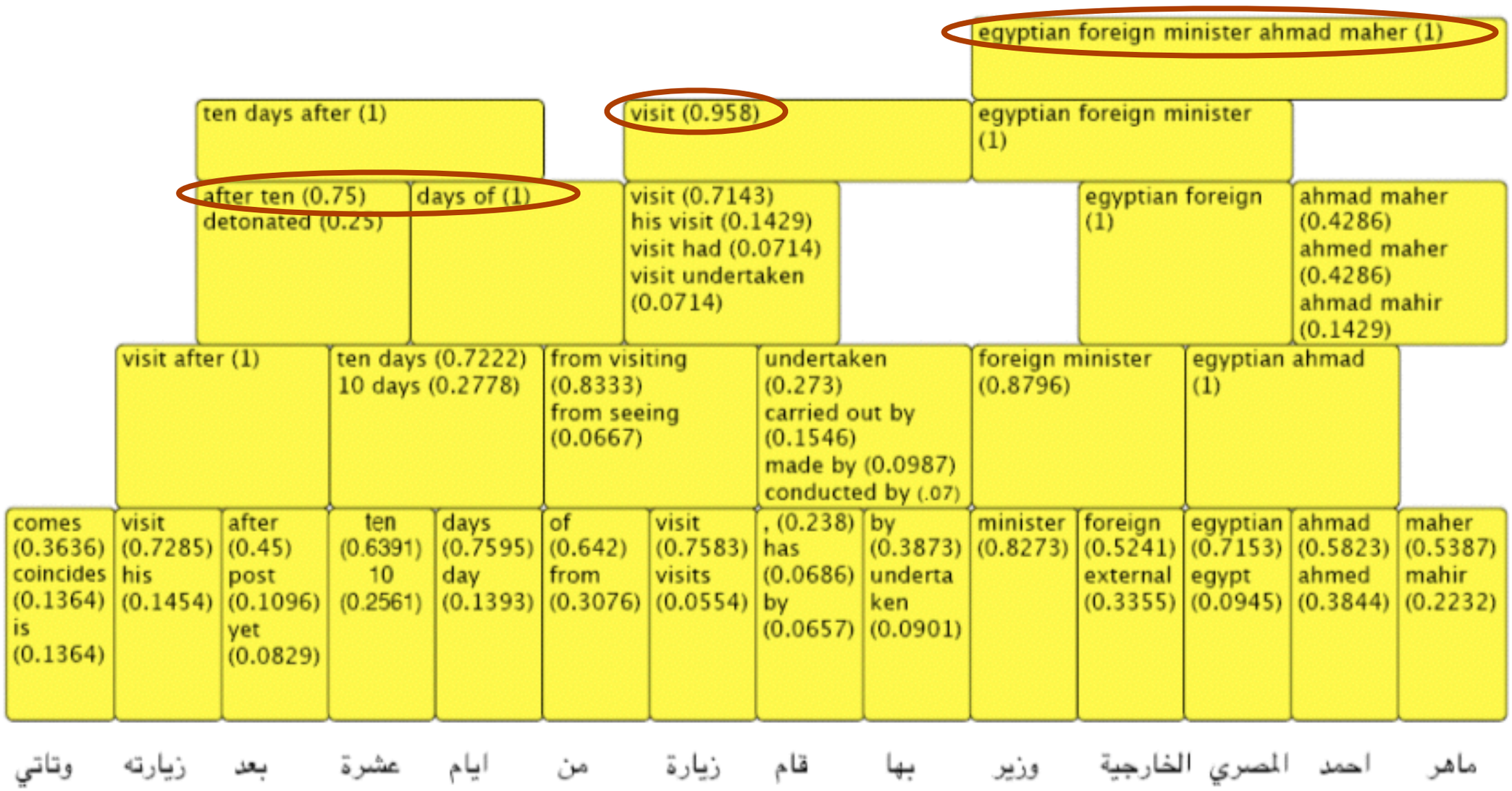
Google (Slav Petrov, SYRCoDIS’11):



Выбираем все консистентные выравнивания

Декодирование

- Аналог Витерби: выбрать предложение e максимизирующее P(e) x P(f | e)



Оценка моделей

- Оценка людьми
  - плавность
  - достоверность
    - адекватность (по фиксированной шкале)
    - информативность (ответ на вопрос по переводу)
- Автоматическая оценка
  - сравнение с одним или несколькими экспертными переводами
  - Меры качества
    - BLEU (Bilingual evaluation understudy)
    - NIST
    - TER
    - METEOR

Оценка моделей: BLEU

- Определить число N-грамм из машинного перевода в экспертных переводах
- Вычислить модифицированную меру точности

Оценка моделей: BLEU

Cand 1: Mary no slap the witch green
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 1 точность по 1-граммам: 5/6

Оценка моделей: BLEU

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 1 точность по 2-граммам: 1/5

Оценка моделей: BLEU

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Для каждой N-граммы счетчик не должен превышать максимального количества этой n-граммы в любом предложении

Cand 2 точность 1-грамм: 7/10

Оценка моделей: BLEU

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 2 точность 2-грамм: 4/9

Модифицированная точность

- Среднее геометрическое всех N-граммам (обычно N < 5)

$$p_n = \frac{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count}\ (\text{n-gram})}$$

$$p = \sqrt[N]{\prod_{n=1}^{N} p_n}$$

Cand 1: $p = \sqrt[2]{\frac{5}{6}\frac{1}{5}} = 0.408$

Cand 2: $p = \sqrt[2]{\frac{7}{10}\frac{4}{9}} = 0.558$

Штраф за краткость

- Сложно посчитать полноту (recall) из-за нескольких экспертных мнений
- Вместо этого используется штраф за краткость
- Пусть *r* - длина экспертного предложения с наибольшим количеством совпадающих N-грамм. Пусть *c* - длина машинного перевода

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Подсчет BLEU

- В итоге: BLEU = BP x p

Cand 1: Mary no slap the witch green.

Best Ref: Mary did not slap the green witch.

$$c = 6, \quad r = 7, \quad BP = e^{(1-7/6)} = 0.846$$

$$BLEU = 0.846 \times 0.408 = 0.345$$

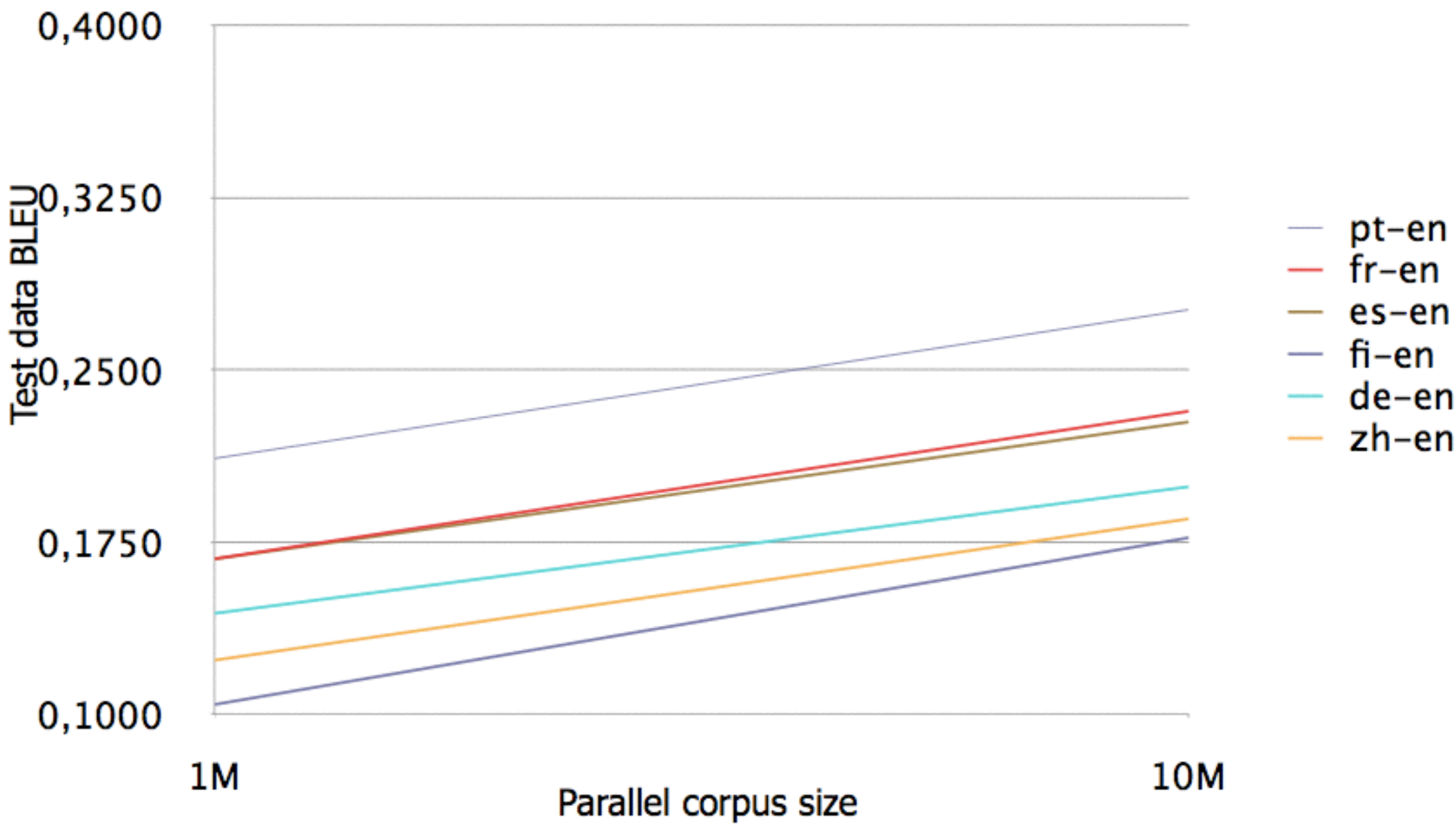
Cand 2: Mary did not give a smack to a green witch.

Best Ref: Mary did not smack the green witch.

$$c = 10, \quad r = 7, \quad BP = 1$$

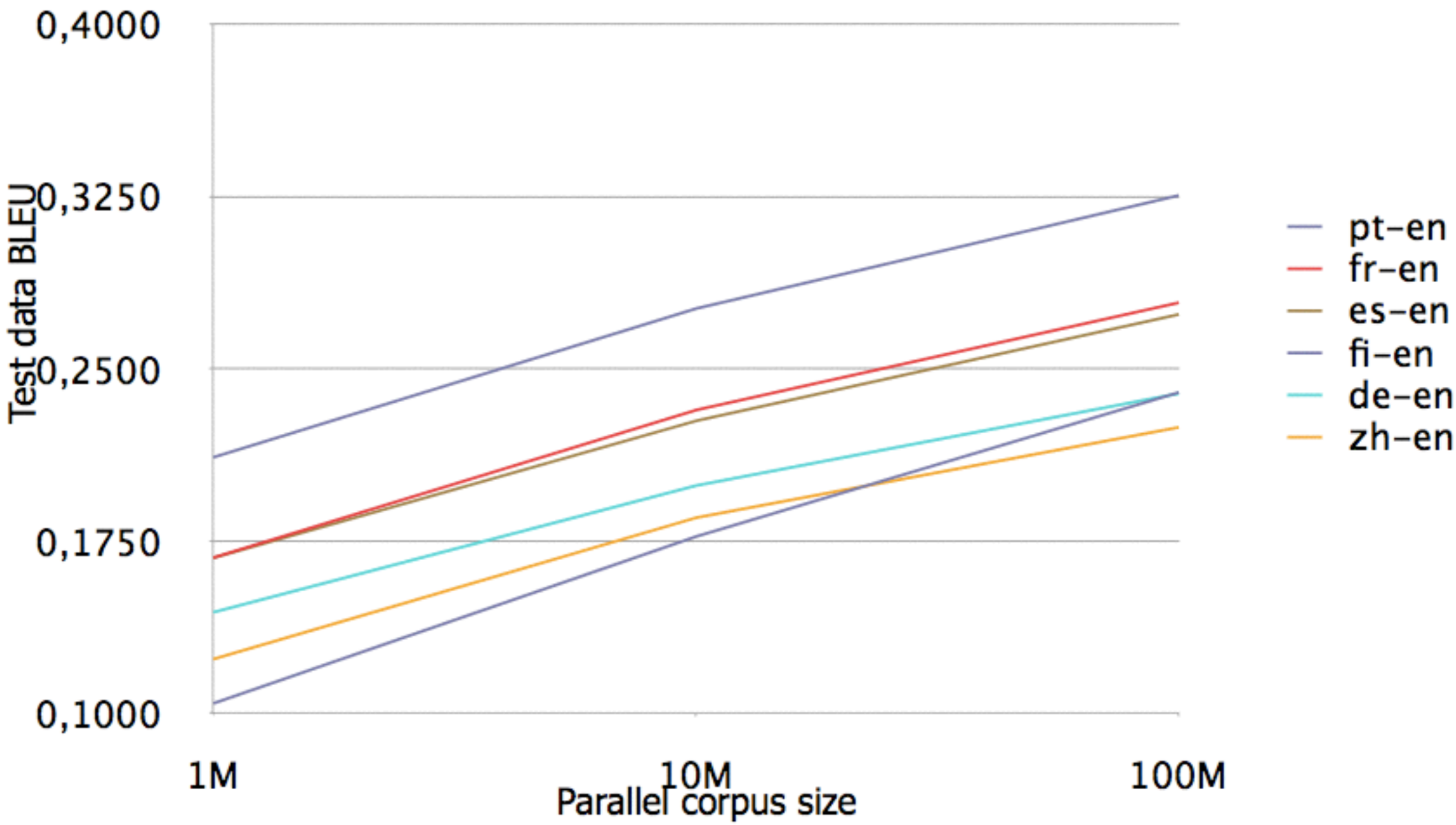
$$BLEU = 1 \times 0.558 = 0.558$$

Лучшие данные - много данных



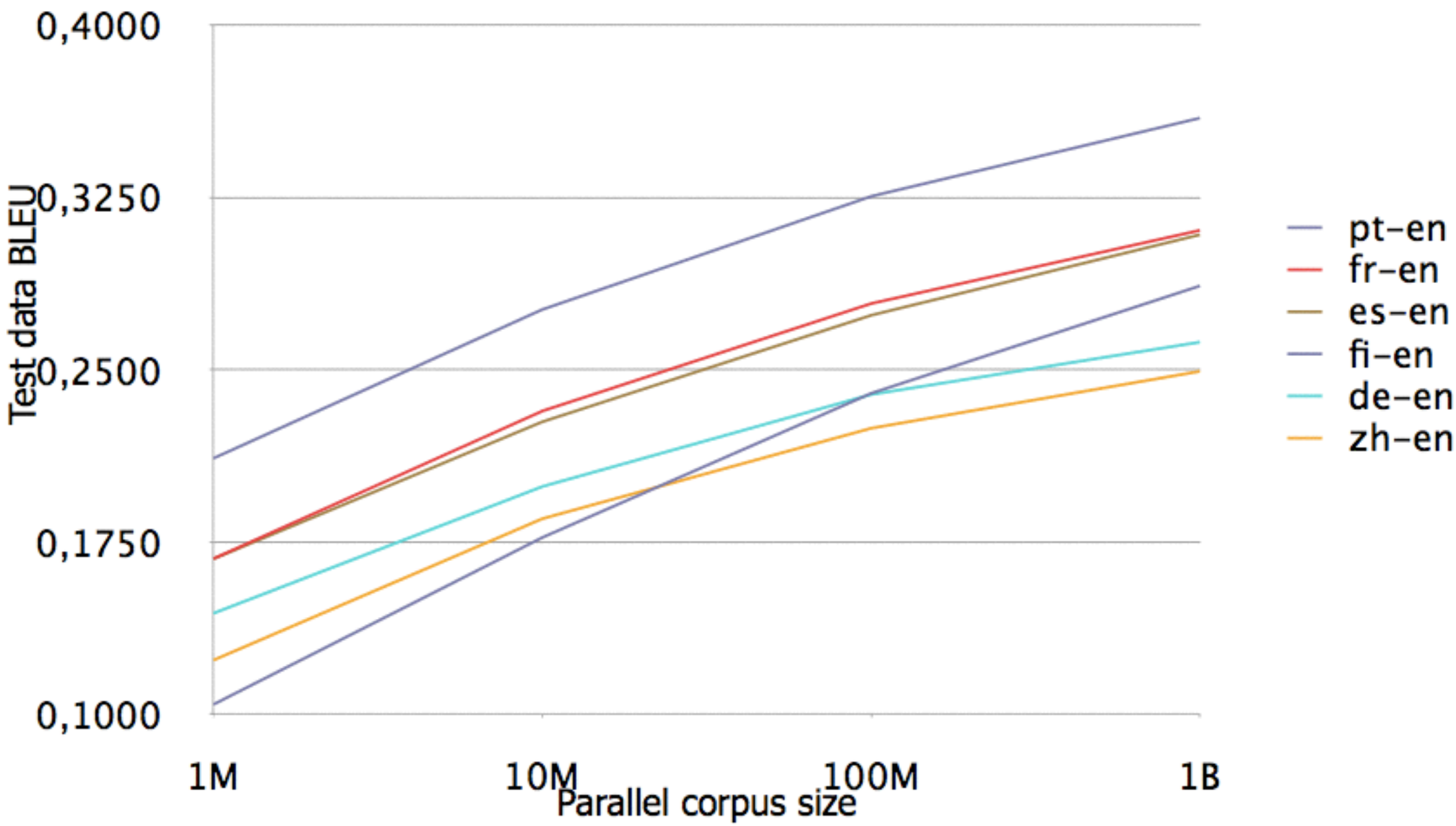
by Google

Лучшие данные - много данных



by Google

Лучшие данные - много данных



by Google

Заключение

- Трудность перевода заключается в существенных **различиях** между языками
- Классические подходы: **прямой перевод, преобразование, интерлингва**
- Для статистического машинного перевода применяется **модель зашумленного канала**, комбинирующая **модель перевода** и **языковую модель**
- Для **выравнивания** слов в двуязычных корпусах применяются формальные модели, например, **IBM Model 1**
- Для оценки систем используются различные метрики: **BLEU**, TER, METEOR.

Следующая лекция

- Тематическое моделирование