Математические модели в лингвистике

Автоматический перевод.

Мати Пентус, Александр Пиперски, Алексей Сорокин

МГУ им. М. В. Ломоносова, межфакультетский курс,
осенний семестр 2017–2018 учебного года, 8 ноября

План лекции

Источник: Statistical machine translation, Philipp Koehn

- Вероятностная модель перевода.
- Математическая модель выравнивания.
- Модель канала связи.
- Языковые модели.
- Восстановление оптимального перевода.
- Использование лингвистической информации при переводе.

Вероятностная модель перевода

- Сколько есть вариантов перевода для Mary has not attended her classes today.
- Варианты перевода:
  - Сегодня Марии не было на занятиях.
  - Сегодня Мария отсутствовала на занятиях.
  - Сегодня Мария пропустила занятия.
  - Сегодня Мария не посетила занятия.
  - Марии не было/Мария отсутствовала на сегодняшних занятиях.
- Кроме того, возможны вариации в зависимости от контекста (Мария/Маша, на занятиях / на уроках / в классе / на лекциях и т.д.).
- Следовательно, нет единственно правильного перевода, есть наиболее вероятный.

Математические обозначения

- Нужен наиболее вероятный перевод: $\widehat{t} = \operatorname{argmax}_t p(t|s)$, где s — исходное предложение, а t — возможный перевод.
- Нам будут полезны следующие формулы:
- Формула произведения вероятностей: $p(x, y) = p(x)p(y)$, если x и y независимы.
- Формула условной вероятности: $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$.
- Формула полной вероятности: если $y_1, \ldots, y_n$ — полное множество взаимоисключающих событий, то

$$p(x) = \sum_i p(x, y_i).$$

- Формула Байеса:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

Мотивация

- Статистическая модель перевода обучается за счёт большого множества параллельных текстов.
- Например, Europarl(корпус заседаний Европарламента), Hansard(французский/английский), EMEA, ...
- При этом неизвестно, какое слово какому соответствует.
- Иногда нет даже соответствия между предложениями, но его легко установить.
- Восстановить выравнивание между словами внутри предложения — отдельная более сложная задача.

Неформальная идея

- При переводе отдельных слов имеет место неоднозначность:

русский	English	p(t|s)
дом	house	0.7
дом	home	0.2
дом	building	0.05
дом	...	...

- Надо как-то извлечь эти вероятности из корпуса.
- Друг другу соответствуют слова из одного предложения:

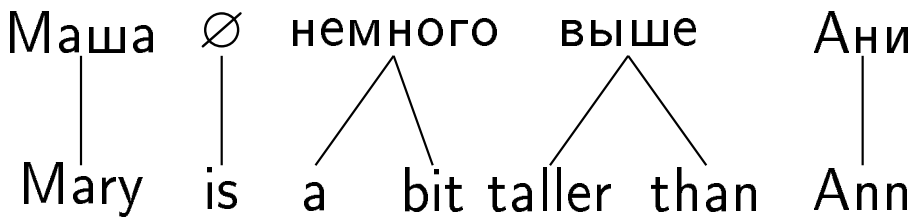
$$\begin{aligned} \text{das Haus} &\leftrightarrow \text{the house}, \\ \text{das Buch} &\leftrightarrow \text{the book}, \\ \text{ein Buch} &\leftrightarrow \text{a book} \end{aligned}$$

- $count(Buch, book) = count(das, the) = 2$,
- $count(Haus, house) = count(ein, a) = 1$.
- Для начала надо установить соответствие между словами.
- То есть построить *выравнивание*..

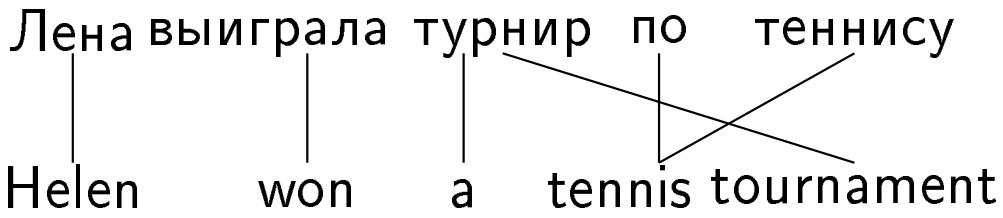
Примеры выравниваний

- Тождественное выравнивание:

- Одно русское слово $\mapsto$ два английских:



- Два русских слова $\mapsto$ одно английское, одно английское $\mapsto$ два русских, изменённый порядок слов:



Вычисление вероятности выравнивания

- Вначале мы не знаем выравнивание, поэтому суммируем вероятность по всем выравниваниям:

$$p(t|s) = \sum_a p(t, a|s)$$

- Как вычислить $p(t, a|s)$?
- Упрощающие предположения:
  - Рёбра выравнивания независимы друг от друга.
  - Каждому слову в s соответствует 1 или 0 слов в t.
- Тогда выравнивание можно задать функцией

$$f : [1, \ldots, |s|] \mapsto [0, \ldots, |t|]$$

.

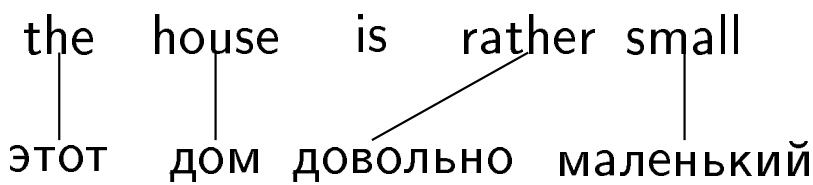
- $f(i) = j \Leftrightarrow s_i$ переводится как t_j.
- $f(i) = 0 \Leftrightarrow s_i$ ничего не соответствует.

Вероятность выравнивания

- Вероятность выравнивания (поменяли направление перевода):

$$P(a, s|t) = \prod_{i=1}^{|s|} \frac{1}{|t|+1} p(s_i|t_{f(i)}) = \frac{1}{(|t|+1)^{|s|}} \prod_{i=1}^{|s|} p(s_i|t_{f(i)})$$

- $\frac{1}{|t|+1}$ — вероятность того, что $f(i) = j$ (равномерное распределение на множестве возможных позиций перевода).
- $p(s_i|t_{f(i)})$ — вероятность слова s_i быть переводом $t_{f(i)}$.
- Пусть $p(house|дом) = 0.7$, $p(is|\varnothing) = 0.1$, $p(the|этот) = 0.2$, $p(rather|довольно) = 0.8$, $p(small|маленький) = 0.4$



- $p(the\ house\ is\ rather\ small, \mathbf{a}|этот\ дом\ довольно\ маленький) = \frac{1}{5^5} 0.7 * 0.1 * 0.2 * 0.8 * 0.4 = 0.0000014336.$

EM-алгоритм: мотивация

- Для вычисления вероятности выравнивания нужны вероятности перевода отдельных слов.
- Но для их извлечения из корпуса нужно выравнивание (какое слово какому соответствует).
- Нельзя ли вычислять обе вероятности одновременно?
  - Вначале задаём равномерное распределение на множестве выравниваний.
  - Слова, часто встречающиеся в одном предложении, будут чаще переходить друг в друга.
  - Отдельные выравнивания станут вероятнее других.
  - После этого вероятности перевода тоже изменятся...
  - В конце концов процесс сойдётся.

Пересчёт вероятностей: EM-алгоритм

- Если $s_i = u$, $t_j = v$, то в части выравниваний между s и t v переходит в u. Это такие a, для которых $a(i) = j$.
- Вероятность таких выравниваний:

$$p(a(i) = j) \sim p(s_i|t_j) = \frac{p(s_i|t_j)}{\sum_k p(s_i|t_k)}$$

- Ожидаемое количество пар (u, v) в выравнивании между s и t:

$$n_s(u) n_t(v) \frac{p(u|v)}{\sum_k p(u|t_k)}$$

- Суммарное количество таких пар по всем предложениям:

$$(u, v) = \sum_{(s,t)} n_s(u) n_t(v) \frac{p(u|v)}{\sum_k p(u|t_k)}$$

- Формула пересчёта вероятностей:

$$p(u|v) \sim c(u, v) = \frac{c(u, v)}{\sum_w c(w, v)}$$

Языковые модели: мотивация

- Иногда выравнивание не позволяет различить переводы: *маленький шаг* $\mapsto$ *small step* или *little step*?
  - small step — 4680000 вхождений,
  - little step — 620000 вхождений.

  Таким образом, $p_t(\text{small step}) \gg p_t(\text{little step})$.

  Как это учесть в модели?

Байесовская модель перевода

$$\widehat{t} = \operatorname{argmax}_t p(t|s) = \operatorname{argmax}_t \frac{p(s|t)p(t)}{p(s)} = \operatorname{argmax}_t p(s|t)p(t)$$

- $p(s|t)$ — оценивает степень соответствия между s и t,
- $p(t)$ — оценивает, насколько вероятен выходной текст.
- Но как мерить $p(t)$?

n-граммная модель текста

- Как оценить $p(w_1 \dots w_m)$?
- Применим формулу условной вероятности:

$$p(w_1 \dots w_m) = p(w_1)p(w_2|w_1)p(w_3|w_1 w_2) \dots p(w_m|w_1 \dots w_{m-1})$$

- n-граммная модель языка: w_m зависит только от $w_1, \dots, w_{m-1}$.
- Униграммная модель $(n = 1)$: $p(w_1 \dots w_m) = p(w_1) \dots p(w_m)$,
- Биграммная модель $(n = 2)$:
  $p(w_1 \dots w_m) = p(w_1)p(w_2|w_1) \dots p(w_m|w_{m-1})$,
- Триграммная модель $(n = 3)$:
  $p(w_1 \dots w_m) = p(w_1)p(w_2|w_1)p(w_3|w_1 w_2) \dots p(w_m|w_{m-2} w_{m-1})$,
- Чем больше n, тем более длинный контекст можно учесть.
- Однако большие n приводят к большей разреженности данных, поэтому брать $n > 3$ нецелесообразно.

Подсчёт вероятностей по корпусу

Как считать $p(w_n|w_1 \dots w_{n-1})$?

$$p(w_n|w_1 \dots w_{n-1}) = \frac{c(w_1 \dots w_n)}{\sum_w c(w_1 \dots w)}$$

я читал		1864			
я читал	книгу	19	$\frac{19}{1864}$	$\approx$	0.010
я читал	газету	3	$\frac{3}{1864}$	$\approx$	0.002
я читал	лекцию	11	$\frac{11}{1864}$	$\approx$	0.006
я читал	доклад	0	$\frac{0}{1864}$	$=$	0?
я читал	инструкцию	0	$\frac{0}{1864}$	$=$	0?

Аддитивное сглаживание

- Можно применить аддитивное сглаживание:

$$p(t_n|t_1 \dots t_{n-1}) = \frac{c(t_1 \dots t_{n-1} t_n) + \alpha}{c(t_1 \dots t_{n-1} \bullet) + \alpha|D|},$$

  где D — словарь (множество возможных униграмм),
  $\alpha > 0$ — сглаживающее слагаемое
- Теперь уже нет нулевых вероятностей. Но как выбирать значение α?
- Маленькая α — риск переподгонки под обучающую выборку.
- Большая α — не учитываем наблюдаемые вероятности.
- α должна зависеть от размера корпуса, словаря и других параметров — слишком сложно подобрать оптимально.
- Кроме того, хотелось бы учитывать сами энграммы.

Интерполяция и откат

- Недостатки аддитивного сглаживания:
  - непонятно, как подбирать α (зависит от размера корпуса, размера словаря, порядка энграмм и т. д.)
  - метод негибкий, не учитывает историю $t_1 \dots t_{n-1}$.
- Основная идея: будем использовать $p(t_n|t_2 \dots t_{n-1})$ для вычисления $p(t_n|t_1 \dots t_{n-1})$, если $c(t_n|t_1 \dots t_{n-1}) = 0$.
- Общая интерполяционная формула:

$$p_I(t_n|t_1 \dots t_{n-1}) = \lambda p_C(t_n|\mathbf{t}_{1,n-1}) + (1-\lambda)p_I(t_n|\mathbf{t}_{2,n-1})$$

$p_C(t_n|t_1 \dots t_{n-1}) = \dfrac{c(t_1 \dots t_{n-1}t_n)}{c(t_1 \dots t_{n-1}\bullet)}$ — "корпусная" вероятность,

λ — коэффициент, вообще говоря, зависящий от $t_1 \dots t_{n-1}$.

Пример

$w_1\,w_2$	w_3	$c(w_1\,w_2\,w_3)$	$p(w_3\|w_1\,w_2)$	w_2	w_3	$c(w_2\,w_3)$	$p(w_3\|w_2)$
я читал		1832		читал		18149	
я читал	газету	3	0.0016	читал	газету	149	0.0082
я читал	книгу	19	0.0103	читал	книгу	138	0.0076
я читал	лекцию	11	0.0060	читал	лекцию	81	0.0045
я читал	доклад	0	0	читал	доклад	22	0.0012

При $\lambda = 0.5$ получаем

$$p(\text{газету}|\text{я читал}) = 0.5 * 0.0016 + 0.5 * 0.0082 = 0.0049$$
$$p(\text{доклад}|\text{я читал}) = 0.5 * 0.0000 + 0.5 * 0.0012 = 0.0006$$

Интерполяция и откат

- Обозначим $\mathbf{t}_{i,j} = t_i \dots t_j$.
- Общая интерполяционная формула:

$$p_I(t_n|\mathbf{t}_{1,n-1}) = \lambda p_C(t_n|\mathbf{t}_{1,n-1}) + (1-\lambda)p_I(t_n|\mathbf{t}_{2,n-1})$$

- Формула отката (backoff):

$$p_I(t_n|t_1 \dots t_{n-1}) = \begin{cases} \alpha(\mathbf{t}_{1,n-1})p_C(t_n|\mathbf{t}_{1,n-1}), & c(\mathbf{t}_{1,n-1}t_n) > 0, \\ \beta(\mathbf{t}_{1,n-1})p_I(t_n|\mathbf{t}_{2,n-1}), & c(\mathbf{t}_{1,n-1}t_n) = 0 \end{cases}$$

- Чем больше λ (α в формуле отката), тем больше мы доверяем истории $\mathbf{t}_{1,n-1}$.
- Много случайных продолжений у $\mathbf{t}_{1,n-1}$ — λ мало.
- Продолжений мало и они частотные — $\lambda \approx 1$
- β подбирают, чтобы сумма вероятностей получилась 1.

Метод Уиттена-Белла

- Метод Уиттена-Белла:

$$
\begin{aligned}
p_I(t_n|\mathbf{t}_{1,n-1}) &= \lambda p_c(t_n|\mathbf{t}_{1,n-1}) + (1-\lambda) p_I(t_n|\mathbf{t}_{2,n-1}) \\
\lambda &= \frac{c(t_1 \dots t_{n-1}\odot)}{c(t_1 \dots t_{n-1}\odot) + N_{1+}(t_1 \dots t_{n-1})} \\
N_{1+}(t_1 \dots t_{n-1}) &= |\{t|c(t_1 \dots t_{n-1} t) > 0\} \\
N_{1+}(t_1 \dots t_{n-1}) &- \text{“число продолжений”}
\end{aligned}
$$

- Пример (британский национальный корпус):

w_1	$c(w_1\odot)$	$N_{1+}(w_1)$	$N_{3+}(w_1)$	$\lambda(w_1)$	$1-\lambda(w_1)$
spite	2899	59	15	$\frac{2899}{2899+59} = 0.980$	0.02
stupid	2898	602	117	$\frac{2898}{2898+602} = 0.828$	0.172

- Униграммная модель для *stupid* в 86 раз более значима, чем для *spite*.

Метод Кнезера-Нея

- Метод Уиттена-Белла учитывает количество возможных правых продолжений.
- Можно учитывать и левые:
  - предшественником слова *York* практически всегда будет слово *New*.
  - соответственно, $p(York|w) \approx 0$ при $w \neq new$.
  - при этом $p_{UNI}(York) = \frac{c(York)}{N}$ достаточно велика.
- В методе Кнезера-Нея униграммная вероятность считается по формуле

$$p_{KN}(w) = \frac{N_{1+}(\bullet w)}{\sum_{w'} N_{1+}(\bullet w')}$$

$$N_{1+}(\bullet w) = |\{w_1 | c(w_1 w) > 0\}|$$ — число левых продолжений

Метод Кнезера-Нея

- Для перераспределения вероятностей на новые слова используется дисконтирование (из всех счётчиков вычитается δ).

$$p_0(t_n|\mathbf{t}_{1,n-1}) = \frac{c(\mathbf{t}_{1,n-1}t_n) - \delta}{c(\mathbf{t}_{1,n-1}\bullet)},\ c(\mathbf{t}_{1,n-1}t_n) > 0$$

- В интерполяционной формуле

$$p_{KN}(t_n|\mathbf{t}_{1,n-1}) = p_0(t_n|\mathbf{t}_{1,n-1}) + \beta(\mathbf{t}_{1,n-1})p_{KN}(t_n|\mathbf{t}_{2,n-1})$$

получаем $\beta = \dfrac{\delta N_{1+}(\mathbf{t}_{1,n-1})}{c(\mathbf{t}_{1,n-1}\bullet)}$ (выведите эту формулу).

- Для униграммных вероятностей — формула с предыдущего слайда.

Метод Кнезера-Нея

- В интерполяционной формуле

$$p_{KN}(t_n|\mathbf{t}_{1,n-1}) = p_0(t_n|\mathbf{t}_{1,n-1}) + \beta(\mathbf{t}_{1,n-1})p_{KN}(t_n|\mathbf{t}_{1,n-2})$$

- Основная проблема: поиск оптимальной δ. В стандартной реализации

$$\delta_1 = 1 - 2Y\frac{N_2}{N_1} \qquad \delta_2 = 1 - 3Y\frac{N_3}{N_2}$$

$$\delta_{\geqslant 3} = 1 - 4Y\frac{N_4}{N_3} \qquad Y = \frac{N_1}{N_1 + 2N_2}$$

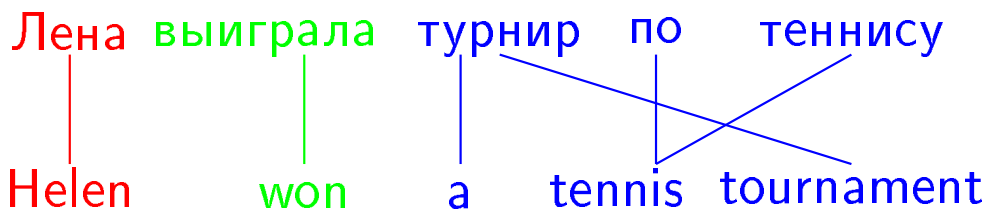
- Здесь δ_i — дисконт для счётчиков, равных i, N_i — число энграмм частоты i.
- В случае лексических энграмм метод Кнезера-Нея наиболее мощный.
- Недостаток: работает только в случае $N_1 \geqslant N_2 \geqslant N_3 \ldots$, поэтому плохо применим к символьным и морфологическим энграммам.

Декодирование

- Мы использовали формулу $\widehat{t} = \text{argmax}_t\, p(s|t)p(t)$.
- Однако невозможно перебрать все возможные переводы, их слишком много. Как найти максимум?
- Первый подход: переводить слово за словом, подбирая локально оптимальный вариант для каждого слова/фразы.
- При локальном выборе руководствуемся лексической вероятностью перевода и языковой моделью.
- При этом гораздо лучше работает перевод по фразам:

Лена	$\leftrightarrow$	Helen
выиграла	$\leftrightarrow$	won
турнир по теннису	$\leftrightarrow$	a tennis tournament

- Тогда эти фразы можно просто соединить в том же порядке:



Фразовый машинный перевод

- Наиболее популярный подход с середины 1990-х гг.
- Переводятся не отдельные слова, а фразы (непрерывные участки текста):

$$\text{No voy}_1 \text{ a la}_2 \text{ casa}_3 \rightarrow \text{I am not going}_1 \text{ to the}_2 \text{ house}_3$$
$$\text{it seems to}_1 \text{ me}_2 \rightarrow \text{me}_1 \text{ parece}_2$$
$$\text{Je}_1 \text{ ne vais pas}_2 \text{ à la}_3 \text{ maison}_4 \rightarrow \text{I}_1 \text{ am not going}_2 \text{ to the}_3 \text{ house}_4$$
$$\text{Eu}_1 \text{ sinto saudade de você}_2 \rightarrow \text{I}_1 \text{ miss you}_2$$
$$\text{I}_1 \text{ miss you}_2 \rightarrow \text{Eu}_1 \text{ sinto sua falta}_2$$
$$\text{natuerlich}_1 \text{ hat}_2 \text{ John}_3 \text{ spass am}_4 \textit{ spiel}_5 \rightarrow \text{of course}_1 \text{ John}_2 \text{ has}_3 \text{ fun with the}_4 \textit{ game}_5$$

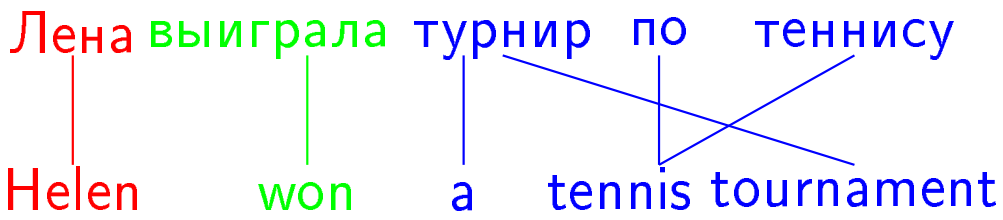
- Тогда вероятность перевода распадается на вероятность перевода по фразам:

$$\begin{aligned} p(s|t) &= \prod_{i=1}^{t} p(s_i|t_i) \\ p(s_i|t_i) &= \frac{}{c(s_i, t_i)} c(t_i) \\ p(s_i|t_i) &- \text{доля } s_i \text{ среди переводов } t_i. \end{aligned}$$

Фразовый машинный перевод

- Локальные изменения порядка слов отражаются внутри фраз.
- На уровне фраз перевод монотонный:

$$\begin{aligned} \text{Лена} &\leftrightarrow \text{Helen} \\ \text{выиграла} &\leftrightarrow \text{won} \\ \text{турнир по теннису} &\leftrightarrow \text{a tennis tournament} \end{aligned}$$



Поиск наилучшего перевода

- Мы искали наилучший перевод по формуле:

$$t^* = \text{argmax}_t \, p(s|t)p(t)$$

- В терминах логарифмов $t^* = \text{argmax}_t \log p(s|t) + \log p(t)$.
- Можно взвешивать вероятности с разными весами:

$$t^* = \text{argmax}_t \, \lambda_1 \log p(s|t) + \lambda_2 \log p(t)$$

- Можно завести некоторое количество признаков $h_i(s, t)$, описывающих перевод из s в t:

$$t^* = \text{argmax}_t \, \lambda_i h_i(s, t)$$

- Дополнительные признаки:
  - Вероятности $p(t|s), p(s|t), p(t)$ (на логарифмической шкале).
  - Разброс $\sum_j \alpha^{start(j+1)-end(j)-1}$ — насколько далеки друг от друга переводы соседних фраз.
  - Количество фраз при переводе.

Поиск наилучшего перевода

- Мы ищем наиболее вероятный / максимизирующий функцию полезности перевод.
- Но невозможно перебрать все переводы: их слишком много.
- Как искать наилучший? Порядок слов может меняться...
- Будем пытаться на каждом шаге перевести по одной фразе исходного предложения и хранить частичный перевод вместе с оценкой его вероятности.
- При этом хранятся только K наилучших гипотез.

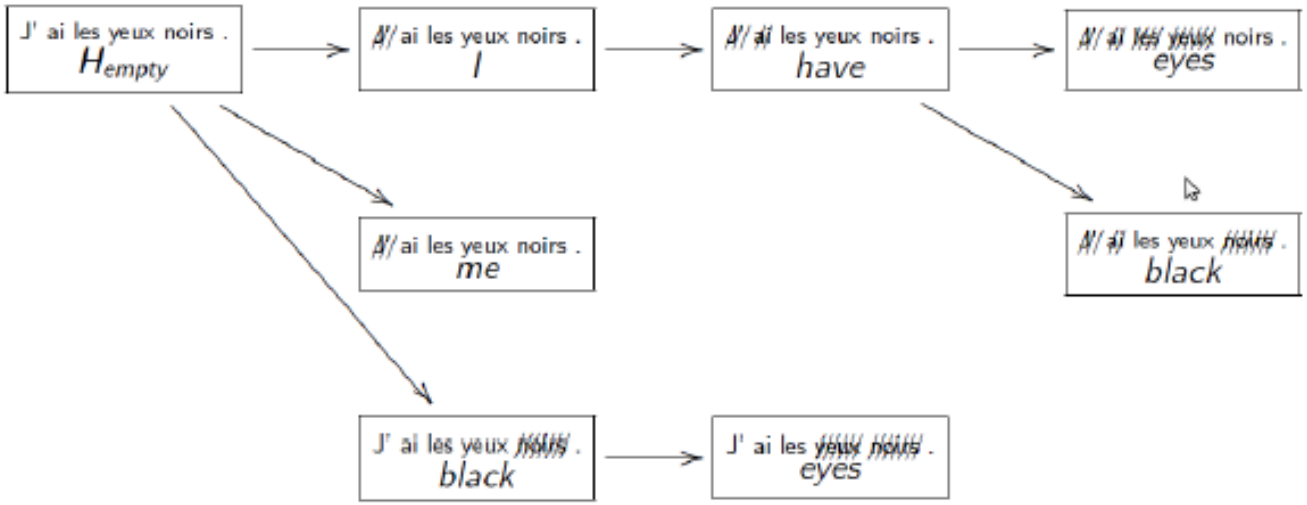
Поиск наилучшего перевода

Пытаемся перебрать все разбиения на фразы:

J'	**ai**	**les**	**yeux**	**noirs**	**.**
I	have	the	eyes	black	.
me	has	them	eye	dark	,
I have		eyes		espresso	!
I am		the eyes		somber	.
I did		some	black eyes		.
I had		black eyes			.
I have		black eyes			.
black eyes		I have			.

Поиск наилучшего перевода

- При переборе храним несколько гипотез и пытаемся их расширить, переводя очередную фразу (необязательно самую левую).



- Гипотеза включает в себя:
  - Переведённые слова.
  - Частичный перевод.
  - Частичную стоимость (с оценкой будущей стоимости).

Привлечение лингвистической информации

- Пока мы никак не привлекали лингвистическую информацию.
- Для похожих и близкородственных языков это работает.
- Однако если в языках разные синтаксические/морфологические системы, одних статистических данных недостаточно.
- Что в одном языке выражено морфологией, в другом выражено синтаксисом:

  ev-ler-imiz-de-ymiş-ler
  house-PL-P1PL-LOC-COP.EV-3PL
  they apparently live in our houses

- Можно рассматривать турецкие суффиксы как отдельные слова.
- Можно отдельно “переводить” морфологические показатели, отдельно — лексемы.

Перевод морфологических показателей

- Можно отдельно “переводить” морфологические показатели, отдельно — лексемы.

$$
\begin{aligned}
\text{выиграла} &= \text{выигрывать+Perf+Past+3+Fem} \\
\text{выигрывать} &\mapsto \text{win} \\
\text{Perf+Past} &\mapsto \text{+Past} \\
\text{+3} &\mapsto \text{+3} \\
\text{+Fem} &\mapsto \varnothing \\
\text{win+Past+3} &= \text{won}
\end{aligned}
$$

- Возникают трудности, если язык-адресат морфологически богаче исходного.
- Mary won $\mapsto$ Маша выиграла/выиграл?
- Нужно делать постобработку с помощью языковой/морфологической модели.