МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №6
по курсу «Автоматическая обработка текстов»

«Применение скрытых марковских моделей к POS-tagging предложения»

Выполнил:
студент группы ИУ9-11М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2018

1 Цель работы

Найти оптимальный морфологический разбор предложения (POS-tagging) «*Time flies like an arrow*», используя скрытую марковскую модель, т.е. вычислить последовательность тэгов, которая максимизирует $ArgmaxP(t_1, t_2, \cdots t_n | w_1, w_2, \cdots w_n)$, где тэги есть части речи.

Исходные данные:

Таблица 1: Матрица A, вероятности перехода между скрытыми состояниями

-	S	Adj	N	V	Conj	Det
S	0	0.2	0.2	0.1	0.01	0.4
Adj	0	0.2	0.5	0.05	0.2	0.01
N	0	0.05	0.01	0.5	0.2	0.01
V	0	0.1	0.2	0.01	0.2	0.3
Conj	0	0.1	0.2	0.2	0	0.2
Det	0	0.3	0.7	0	0	0

Вероятности наблюдаемых состояний:

- $P(time|N) = 0.01, P(time|V) = 0.001, P(time|Adj) = 0.0005$
- $P(flies|N) = 0.0005, P(flies|V) = 0.01$
- $P(like|N) = 0.001, P(like|V) = 0.02, P(like|Conj) = 0.05$
- $P(an|Det) = 0.1$
- $P(arrow|N) = 0.01, P(arrow|Adj) = 0.01$

2 Вычисления

Имеем последовательность наблюдаемых состояний. Для вычисления наиболее вероятной последовательности скрытых состояний воспользуемся алгоритмом Витерби.

2.0.1 Time

Рассчитаем вероятность перехода из начального состояния S в возможные скрытые состояния слова $Time$:

- $P(time|Adj)P(S \to Adj) = 0.0005 * 0.2 = 0.0001$
- $P(time|N)P(S \to N) = 0.01 * 0.2 = 0.002$
- $P(time|V)P(S \to V) = 0.001 * 0.1 = 0.0001$

На данный момент мы не имеем достаточно информации, для того, чтобы выбрать наиболее подходящий тэг для $Time$, так что считаем все переходы равносильными.

2.0.2 flies

Вычислим, переходы из каких состояний (Adj, N или V) в состояния N и V слова $flies$ показывают лучший результат. Для этого умножим получившийся результат на прошлом шаге для состояния X (P_{prev_x}) с вероятностью сделать переход $X \to Y$ и с веротяностью встретить слово $flies$ в состоянии Y, где $X \in \{Adj, N, V\}$, $Y \in \{N, V\}$:

Переход в N:

- $P_{prev\_Adj} * P(Adj \to N) * P(flies|N) = 0.0001 * 0.5 * 0.0005 = 2.5 * 10^{-8}$ – **лучший**
- $P_{prev\_N} * P(N \to N) * P(flies|N) = 0.002 * 0.01 * 0.0005 = 10^{-8}$
- $P_{prev\_V} * P(V \to N) * P(flies|N) = 0.0001 * 0.2 * 0.0005 = 10^{-8}$

Переход в V:

- $P_{prev\_Adj} * P(Adj \to V) * P(flies|V) = 0.0001 * 0.05 * 0.01 = 5 * 10^{-8}$
- $P_{prev\_N} * P(N \to V) * P(flies|V) = 0.002 * 0.5 * 0.01 = 10^{-5}$ – **лучший**
- $P_{prev\_V} * P(V \to V) * P(flies|V) = 0.0001 * 0.01 * 0.01 = 10^{-8}$

2.0.3 like

Аналогичным образом рассчитаем переходы для слова $like$:

Переход в $Conj$:

- $P_{prev\_N} * P(N \to Conj) * P(like|Conj) = 2.5 * 10^{-8} * 0.2 * 0.05 = 2.5 * 10^{-10}$
- $P_{prev\_V} * P(V \to Conj) * P(like|Conj) = 10^{-5} * 0.2 * 0.05 = 10^{-7}$ – **лучший**

Переход в N:

- $P_{prev\_N} * P(N \to N) * P(like|N) = 2.5 * 10^{-8} * 0.01 * 0.001 = 2.5 * 10^{-13}$
- $P_{prev\_V} * P(V \to N) * P(like|N) = 10^{-5} * 0.2 * 0.001 = 2 * 10^{-9}$ – **лучший**

Переход в V:

- $P_{prev\_N} * P(N \to V) * P(like|V) = 2.5 * 10^{-8} * 0.5 * 0.02 = 2.5 * 10^{-10}$
- $P_{prev\_V} * P(V \to V) * P(like|V) = 10^{-5} * 0.01 * 0.02 = 2 * 10^{-9}$ – **лучший**

Таким образом, все переходы в состояния слова $like$ происходят из состояния V. Путь $S \to Adj \to N$ завершился.

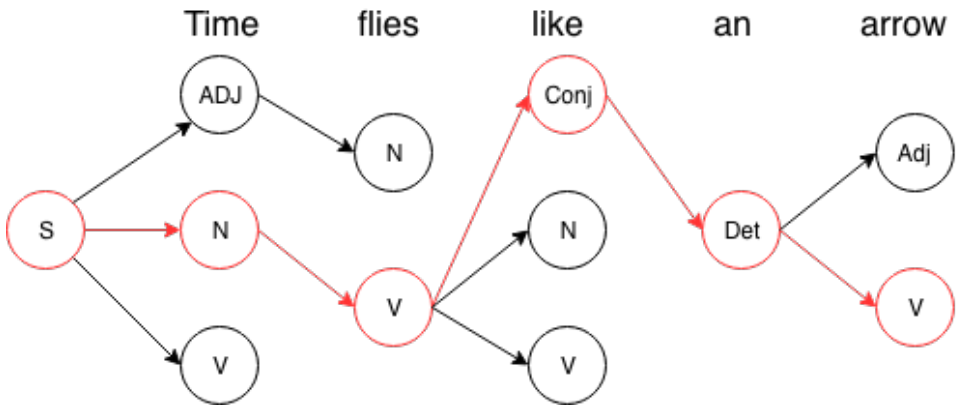

Рис. 1:

2.0.4 an

Рассчитаем наилучший переход в состояние Det слова an и завершим 2 из трех путей.

Переход в Det:

- $P_{prev\_Conj} * P(Conj \to Det) * P(an|Det) = 10^{-7} * 0.2 * 0.1 = 2 * 10^{-9}$ – **лучший**
- $P_{prev\_N} * P(N \to Det) * P(an|Det) = 2 * 10^{-9} * 0.01 * 0.1 = 2 * 10^{-12}$
- $P_{prev\_V} * P(V \to Det) * P(an|Det) = 2 * 10^{-9} * 0.3 * 0.1 = 6 * 10^{-11}$

Таким образом, пути $S \to N \to V \to V$ и $S \to N \to V \to N$ завершились.

2.0.5 arrow

Рассчитаем последний переход.

Переход в Adj:

- $P_{prev\_Det} * P(Det \to Adj) * P(arrow|Adj) = 2 * 10^{-9} * 0.3 * 0.01 = 6 * 10^{-12}$

Единственный переход: $Det \to Adj : 6 * 10^{-12}$

Переход в N:

- $P_{prev\_Det} * P(Det \to N) * P(arrow|N) = 2 * 10^{-9} * 0.7 * 0.01 = 1.4 * 10^{-11}$

Единственный переход: $Det \to N : 1.4 * 10^{-11}$

Теперь, когда построены все пути, необходимо выбрать наилучший (из двух оставшихся). Лучший путь заканчивается состоянием N, т.к. его значение является «наибольшим». Схематично путь представлен на Рисунке 1

3 Выводы

В ходе работы был изучен и осуществлен POS-tagging заданного предложения. Получившийся в ходе работы разбор совпадает с наиболее употребимым в повседневной речи «разбором» этого предложения.