Лингвистика для математиков

POS-tagging

План на сегодня

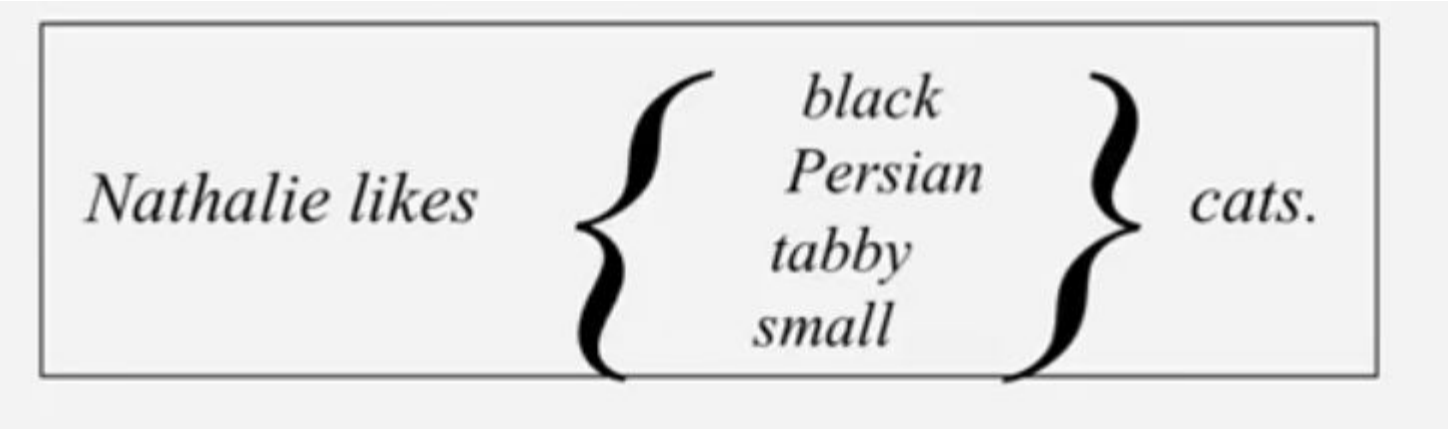
- Автоматическое выделение частей речи
- Пробный тест по фану без оценок

I'M SO ADJECTIVE,
I VERB NOUNS!!!

Какие бывают части речи?

Части речи

Как определить часть речи?



Части речи

Открытые и закрытые

Что это значит?

- Open class:
  - nouns, non-modal verbs, adjectives, adverbs
- Closed class:
  - prepositions, modal verbs, conjunctions, particles, determiners, pronouns

Глокая куздра штеко будланула бокра и кудрячит бокрёнка

Части речи

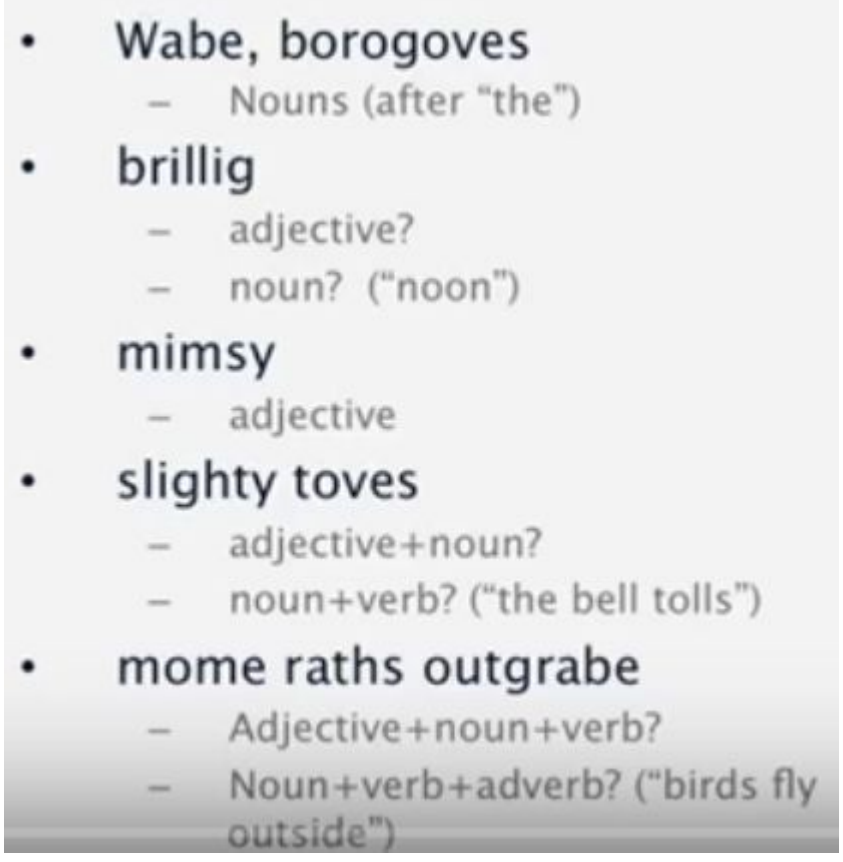
Из Алисы в стране чудес

Lewis Carroll

`Twas **brillig**, and the **slithy toves**
Did gyre and gimble in the **wabe**:
All **mimsy** were the borogoves,
And the **mome raths outgrabe**.

Части речи

Ответы на задачку



Части речи в разных языках

Вспомним задачу про индонезийский

Части речи в русском

Неоднозначность

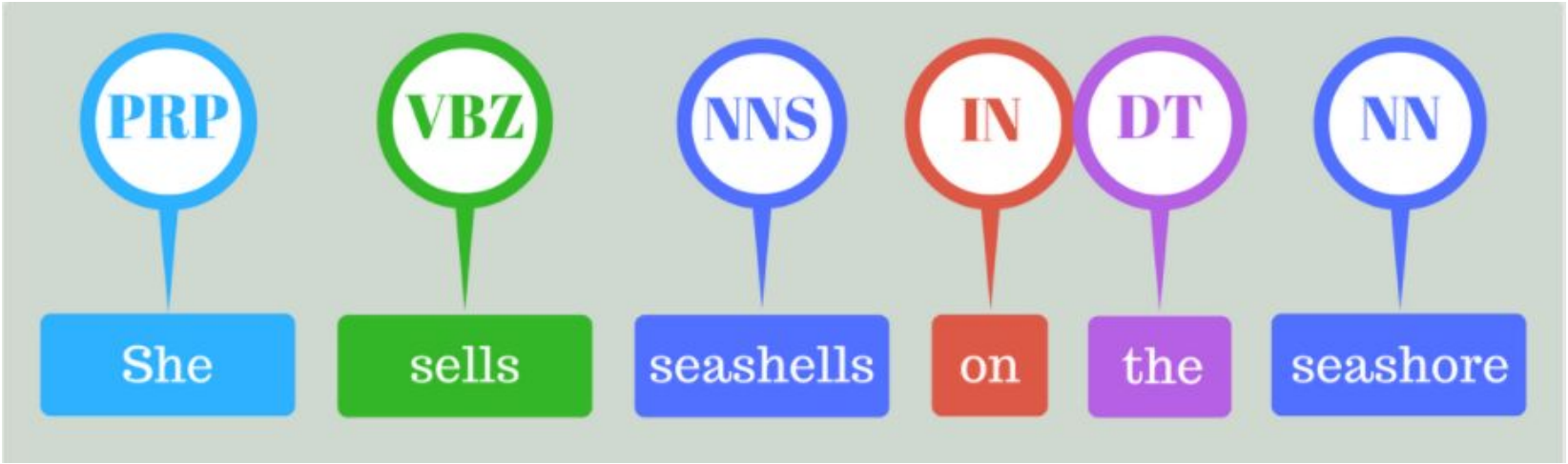
(в английском)

- count (noun) vs. count (verb)
- 11% of all types but 40% of all tokens in the Brown corpus are ambiguous.
- Examples
  - *like* can be tagged as ADP VERB ADJ ADV NOUN
  - *present* can be tagged as ADJ NOUN VERB ADV

Автоматический морфологический анализ

Как автоматически отличить “book that flight” от “hand me this book”?

Нужно провести морфологический анализ



The Penn Treebank tagset

- Университет Пенсильвании. Использовался для ручной разметки корпуса для текстов.

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	*$*
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	*#*
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	“	left quote	‘ or “
LS	list item marker	*1, 2, One*	TO	“to”	*to*	”	right quote	’ or ”
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Figure 8.1 Penn Treebank part-of-speech tags (including punctuation).

Universal dependencies

- Этот набор тегов используется в большинстве современных корпусов
- Используется для большого количества языков
- Можно сравнивать разные языки и делать разборы более однообразными
- + синтаксический парсинг

Точность

- базовый алгоритм: если слово неоднозначно, присваиваем ему ту часть речи, которая чаще всего встречается в корпусе (для этого слова) --- 90% точность
- более сложные алгоритмы (скрытые марковские модели, машинное обучение и т.д.) --- 97% точность
- человек --- 98% точность

Какими методами мы можем воспользоваться?

- на основе сета правил
- стохастические (с помощью машинного обучения, с помощью марковских моделей)

Первый метод: сверяемся с таблицей

	N	V
Mary	1	0
saw	0	2
Jane	2	0
Will	1	0

Mary saw Will.

Data:

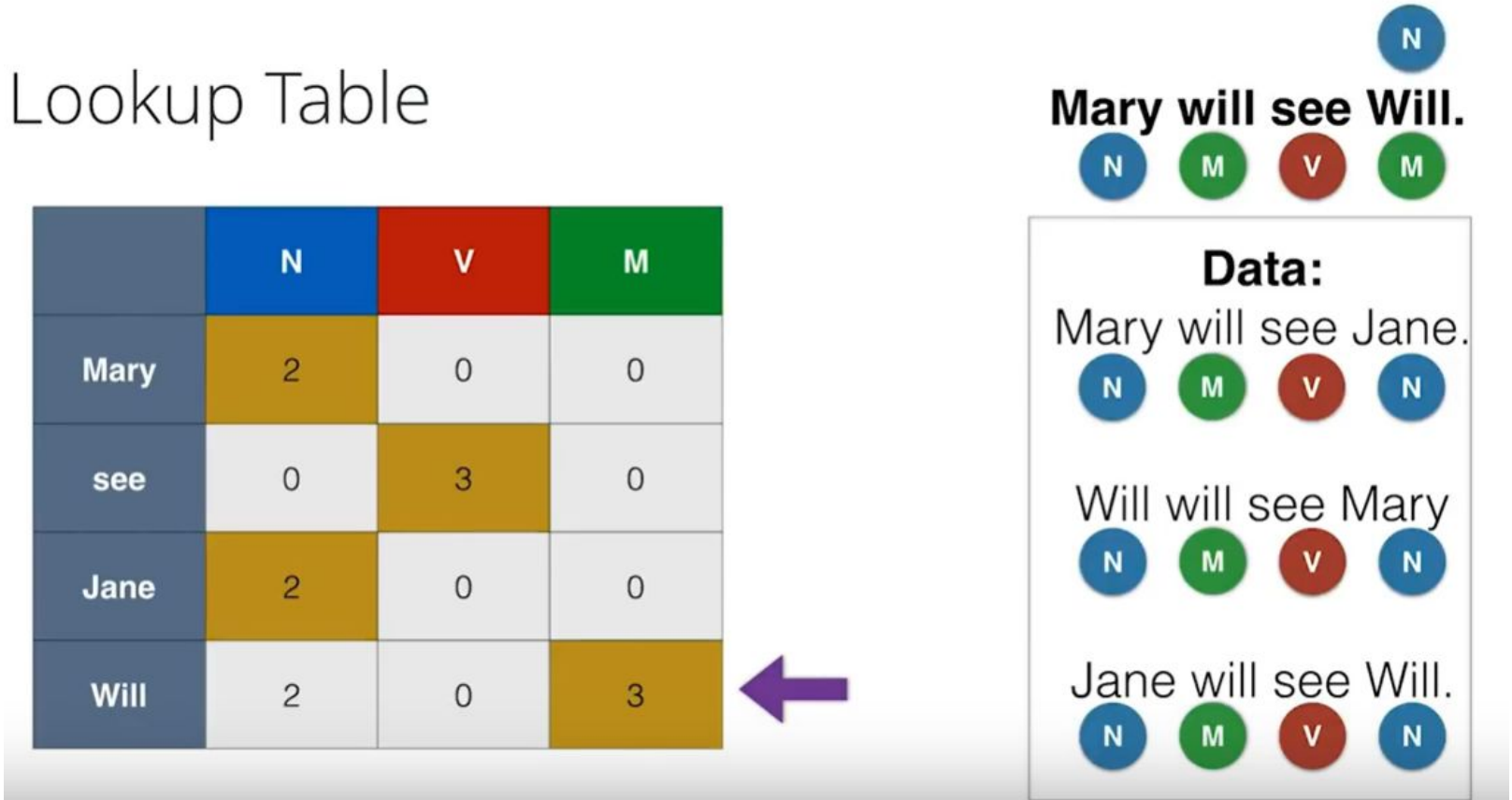
Mary saw Jane.

N V N

Jane saw Will.

N V N

Первый метод: сверяемся с таблицей



Второй метод: n-граммы

Bigrams

	N-M	M-V	V-N
mary-will	1	0	0
will-see	0	3	0
see-jane	0	0	1
will-will	1	0	0
see-mary	0	0	1
jane-will	1	0	0
see-will	0	0	1

Mary will see Will.

N M V N

Data:

Mary will see Jane.

N M V N

Will will see Mary

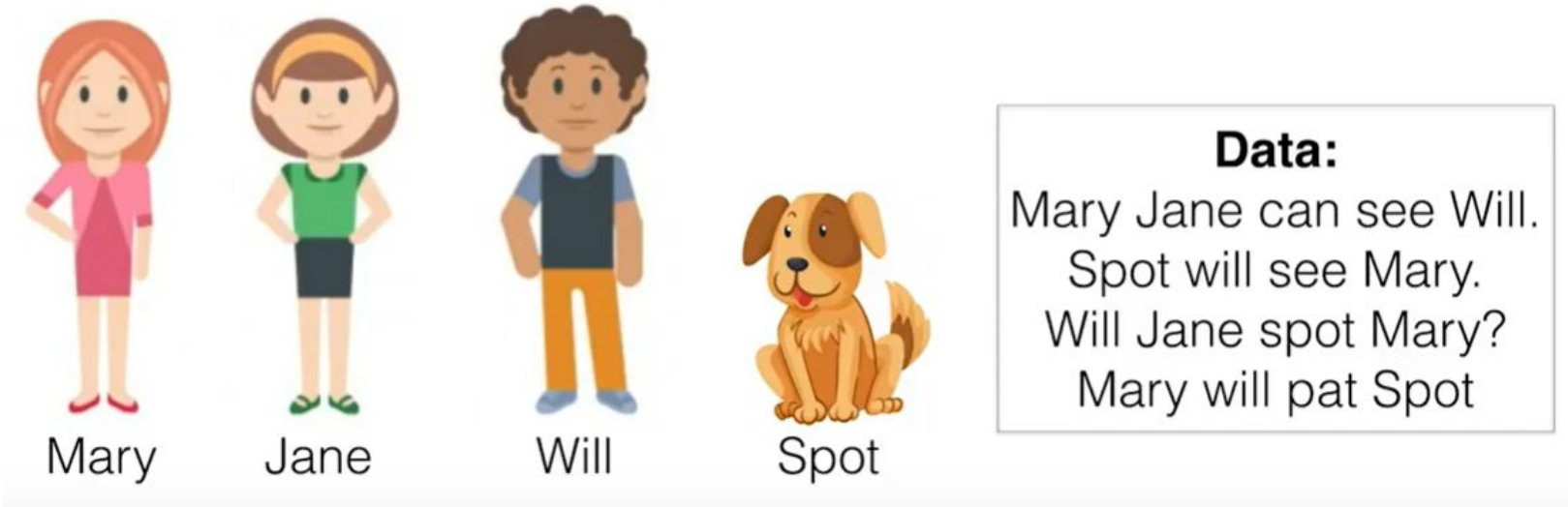
N M V N

Jane will see Will.

N M V N

Второй метод: n-граммы

Jane will spot Mary --- эта пара (биграмм) не встретится в таблице. Как мы тогда присвоим ему частотность/вероятность?



Скрытые марковские модели

Будущее зависит от прошлого только через настоящее

Это называется марковской цепью

Скрытые марковские модели

Сначала классический пример про погоду и настроение

https://www.youtube.com/watch?v=kqSzLo9fenk

до 11 минуты

Скрытые марковские модели

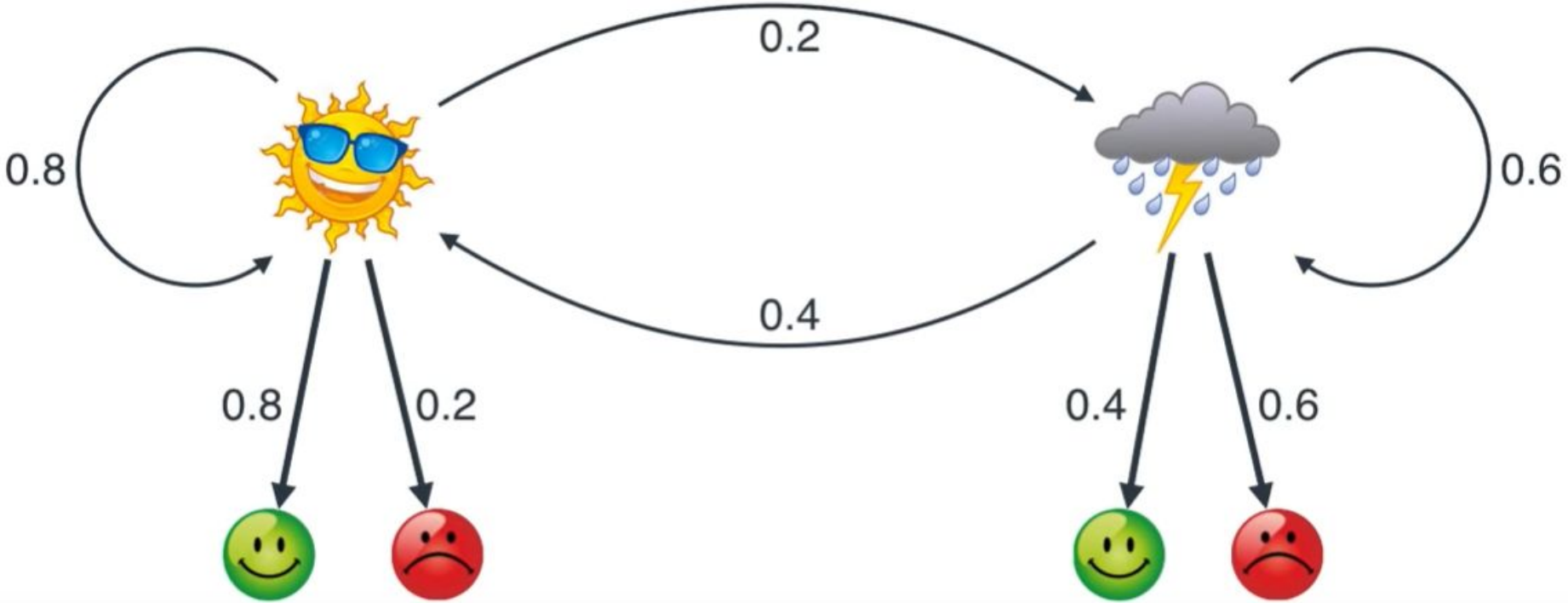
Нам нужна последовательность наблюдений. Событий и каких-то зависимых от них событий

Два типа вероятностей:

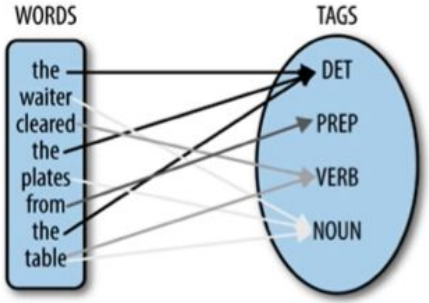
- вероятность перехода из одного состояния в другое
- вероятность того, что при условии, что есть одно состояние, то ему соответствует какое-то событие

Наша первая марковская модель

Как это соотноситься с языком?



Применения скрытых марковских моделей



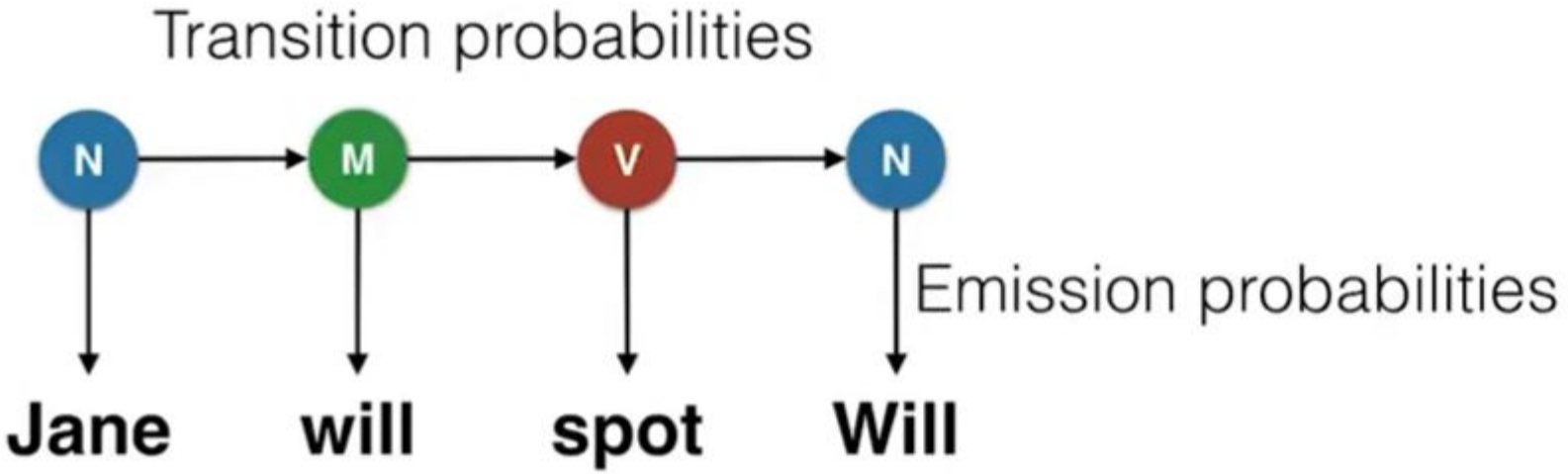
Part of Speech Tagging

Robot Localization

Genetics

Speech Recognition

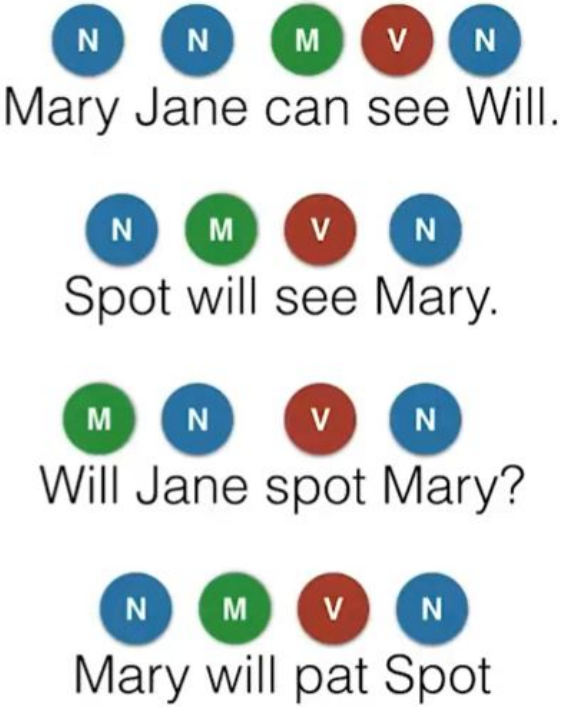
Скрытые марковские модели



Скрытые марковские модели

Emission Probabilities

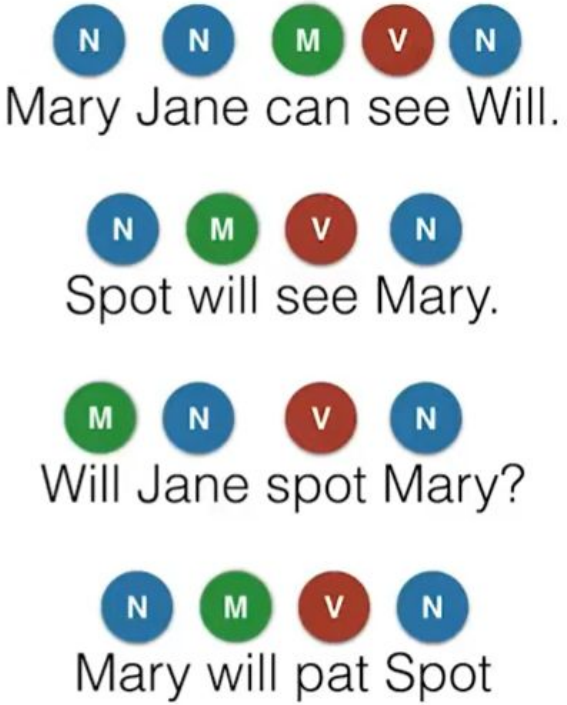
	N	M	V
Mary	4	0	0
Jane	2	0	0
Will	1	3	0
Spot	2	0	1
Can	0	1	0
See	0	0	2
Pat	0	0	1



Скрытые марковские модели

Emission Probabilities

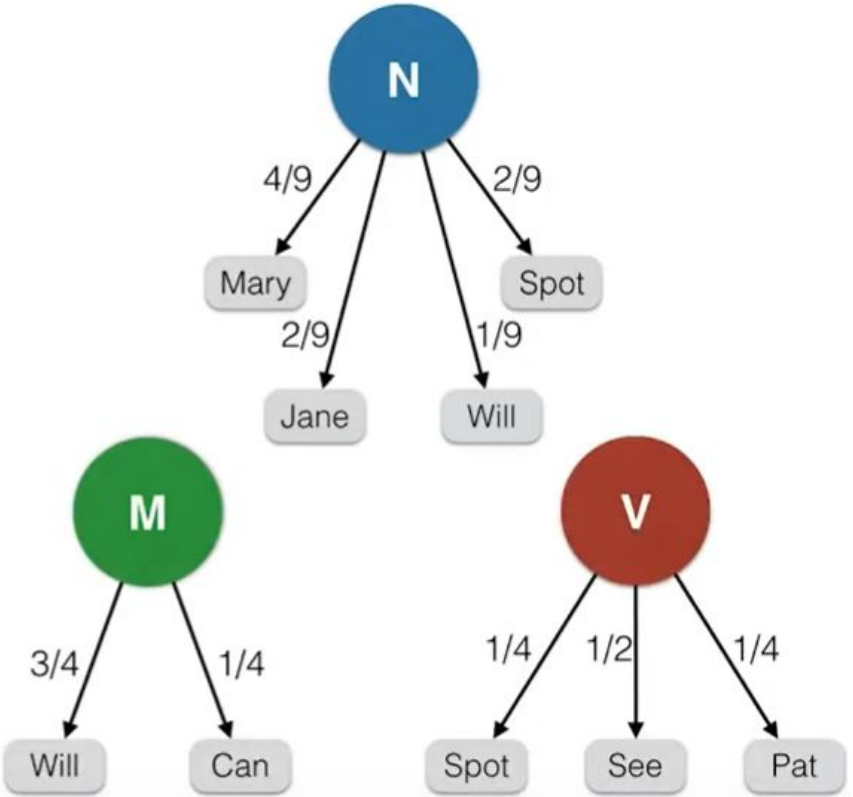
	N	M	V
Mary	4/9	0	0
Jane	2/9	0	0
Will	1/9	3/4	0
Spot	2/9	0	1/4
Can	0	1/4	0
See	0	0	1/2
Pat	0	0	1/4



Скрытые марковские модели

Emission Probabilities

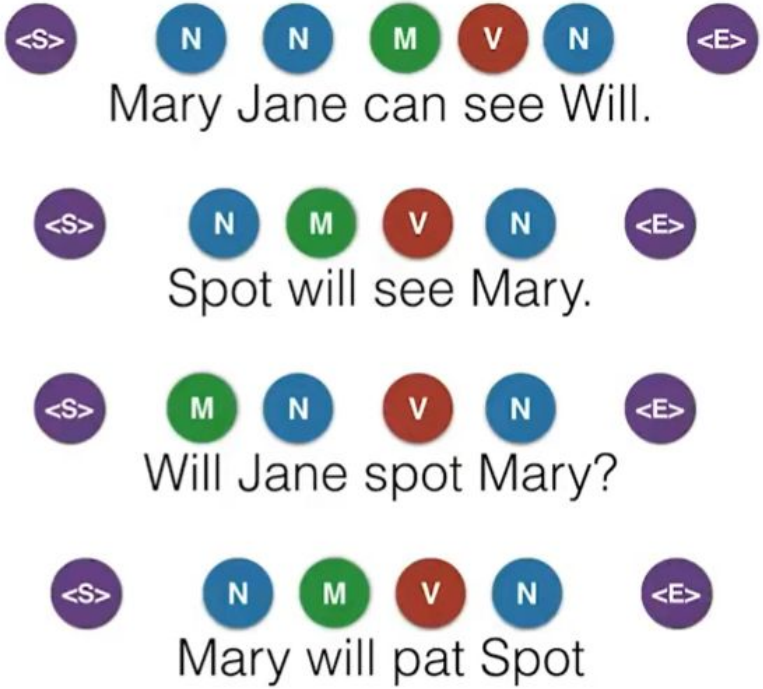
	N	M	V
Mary	4/9	0	0
Jane	2/9	0	0
Will	1/9	3/4	0
Spot	2/9	0	1/4
Can	0	1/4	0
See	0	0	1/2
Pat	0	0	1/4



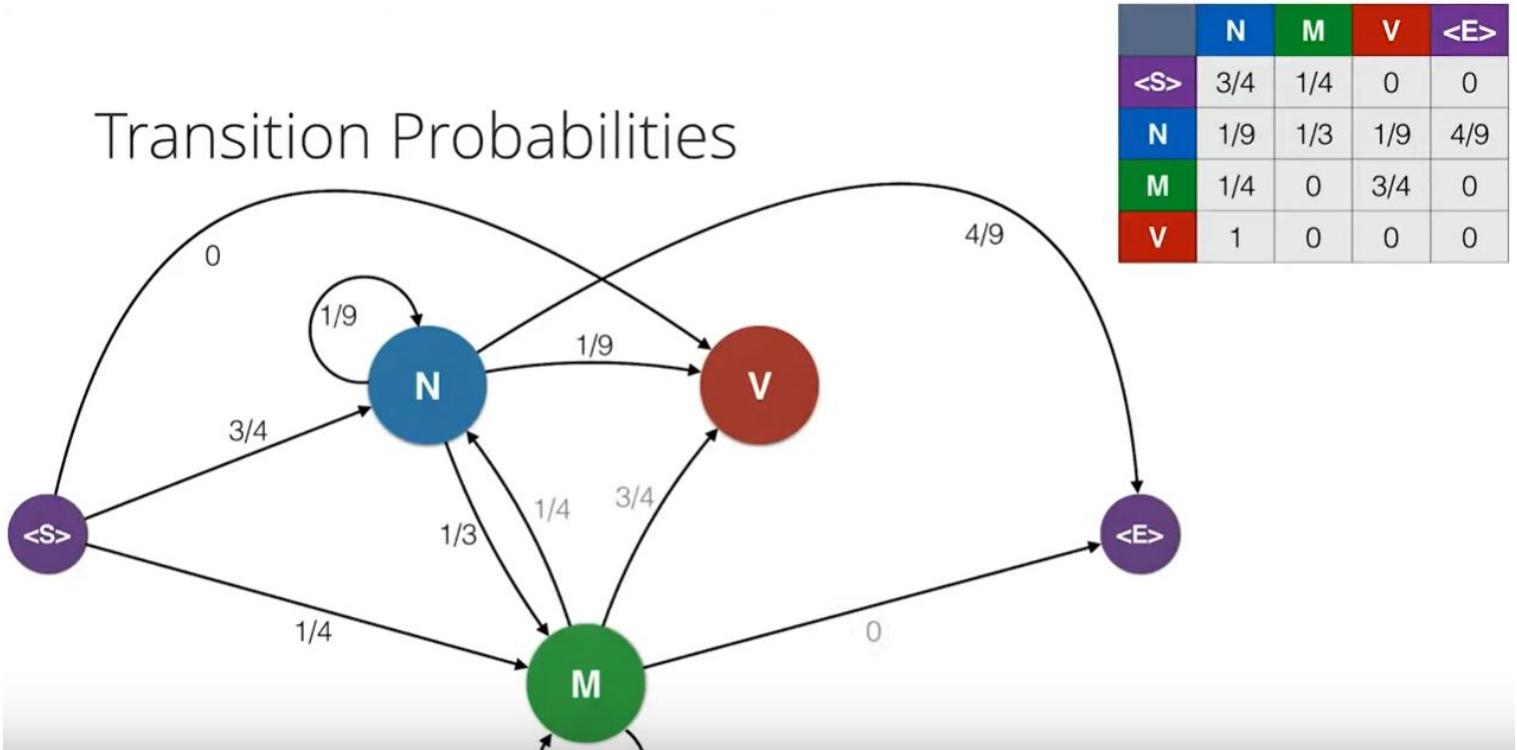
Скрытые марковские модели

Transition Probabilities

	N	M	V	<E>
<S>	3/4	1/4	0	0
N	1/9	1/3	1/9	4/9
M	1/4	0	3/4	0
V	1	0	0	0

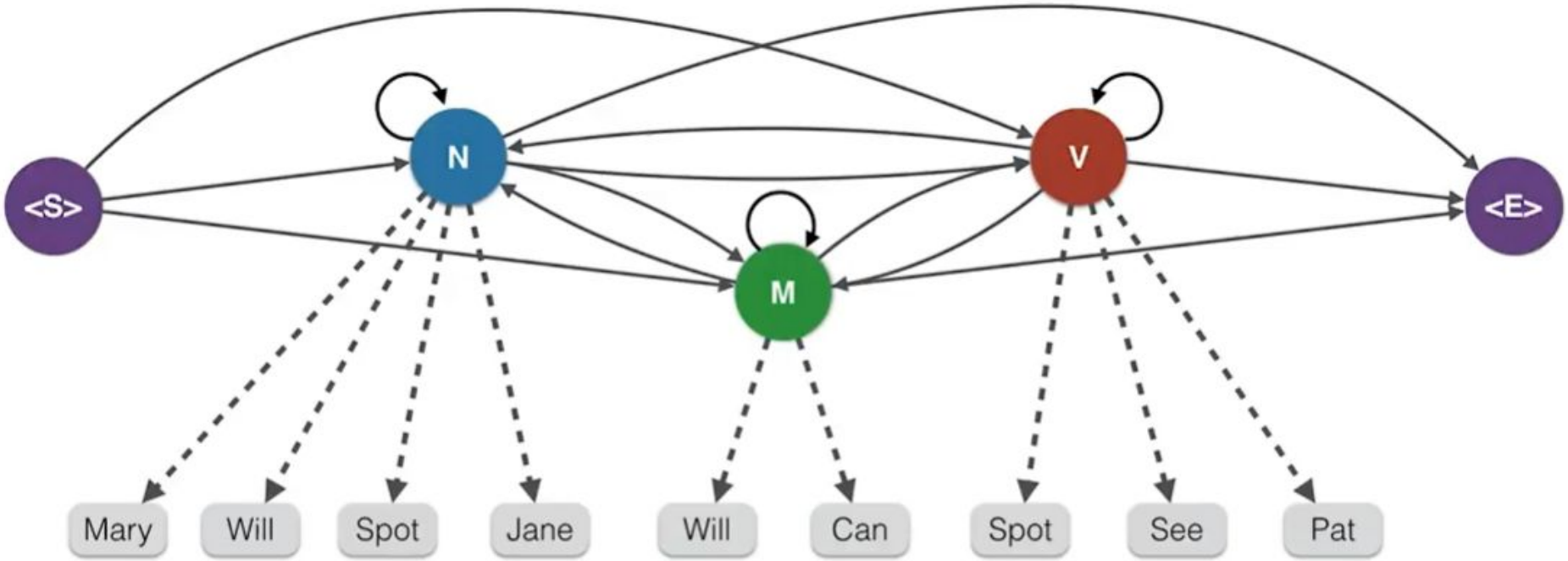


Скрытые марковские модели



	N	M	V	<E>
<S>	3/4	1/4	0	0
N	1/9	1/3	1/9	4/9
M	1/4	0	3/4	0
V	1	0	0	0

Скрытые марковские модели



Скрытые марковские модели

https://www.youtube.com/watch?v=ZDXlExZlVMs&list=PLC0PzjY99Q_U5bba7gYJicCxIufrFmlTa&index=7

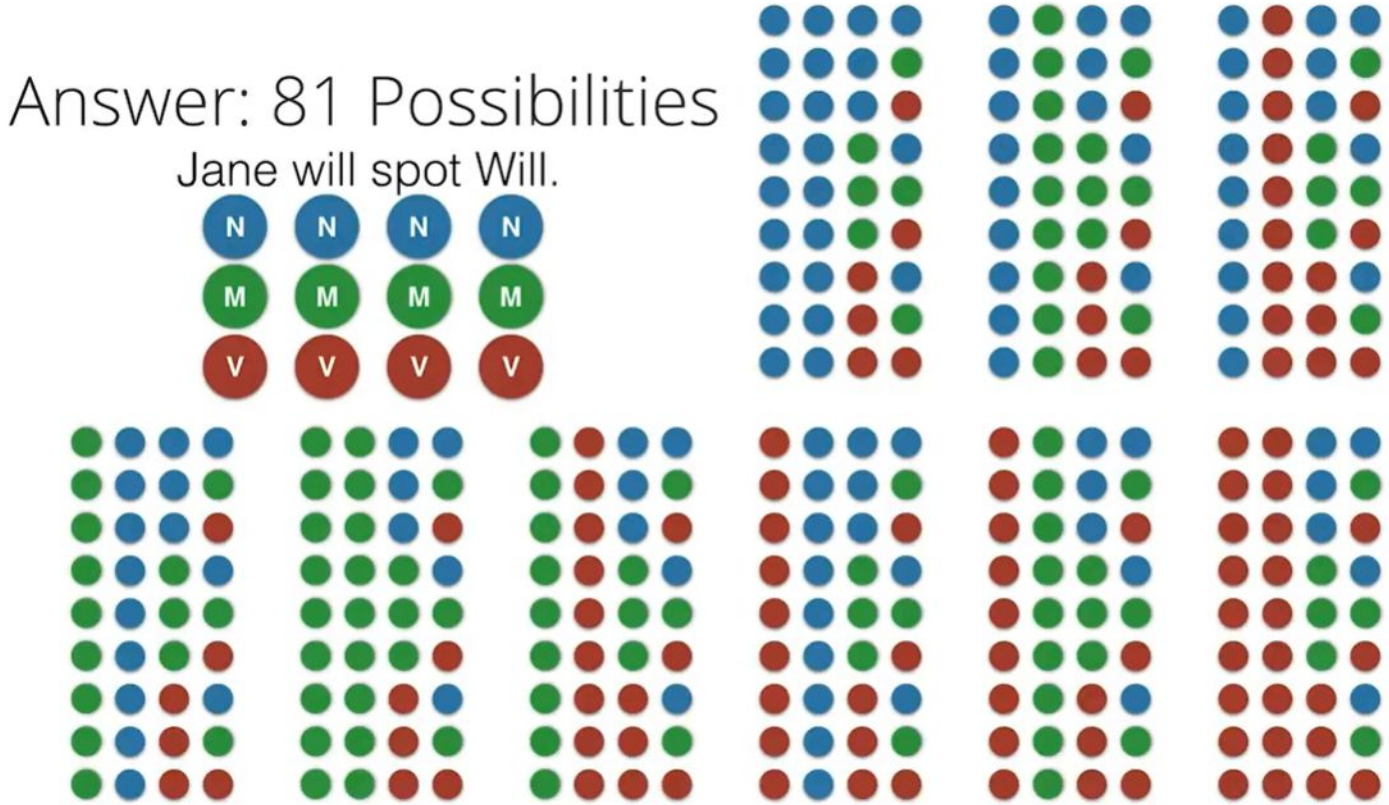
Скрытые марковские модели

Задача:

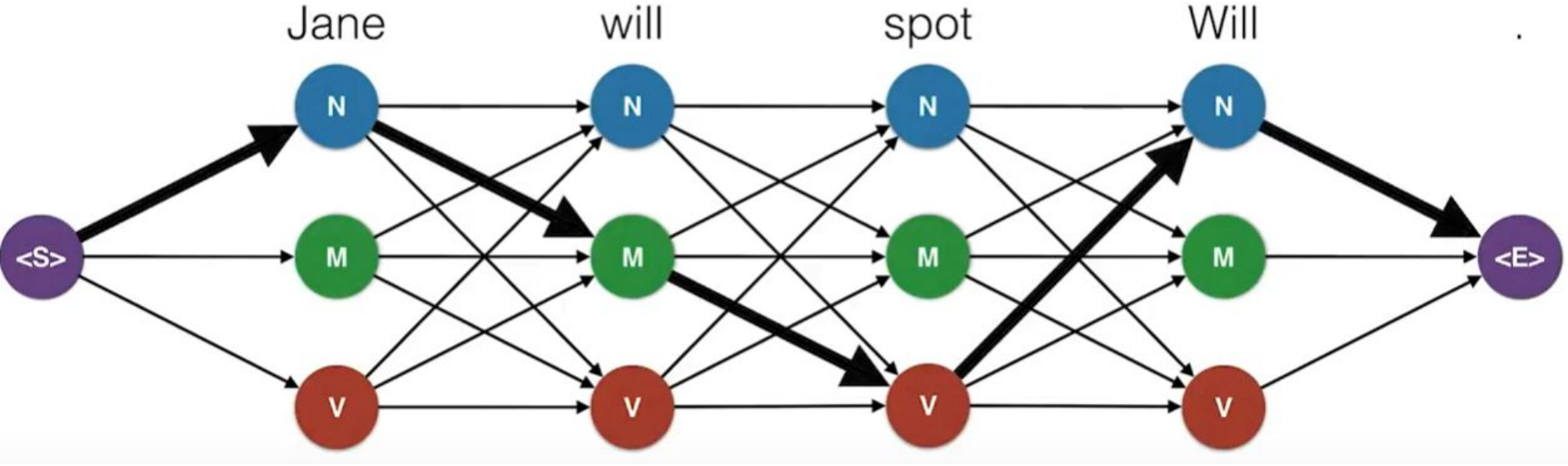
У нас есть 3 части речи: modal verb, verb, noun. Сколько возможных цепочек частей речи нужно проверить скрытой марковской модели для выбора наиболее вероятной для предложения

Jane will spot Will

Скрытые марковские модели



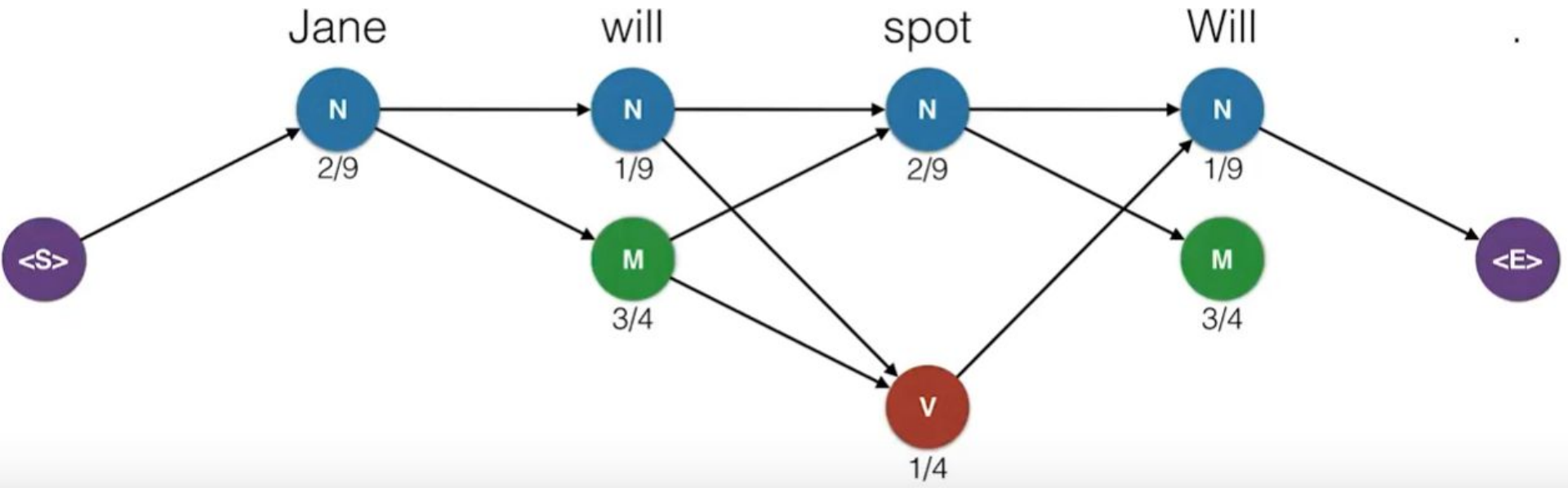
Скрытые марковские модели



Скрытые марковские модели

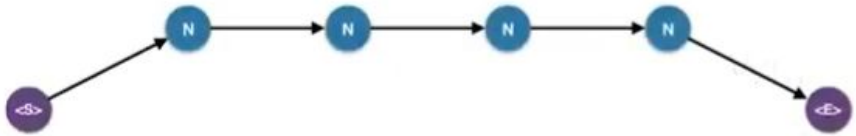
Сколько путей нам нужно проверить теперь?

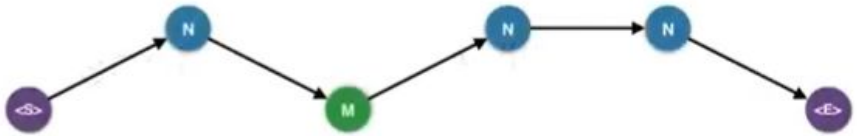
Что мы удалили?

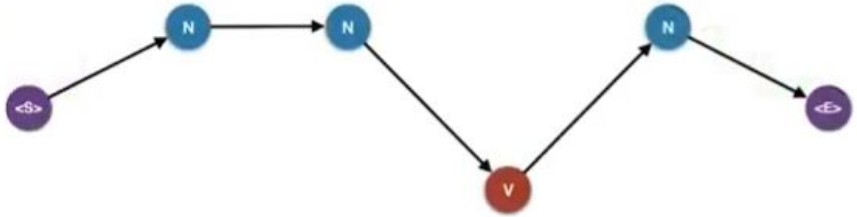


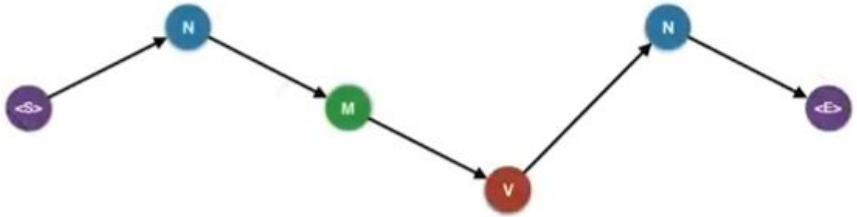
Скрытые марковские модели

Ответ: 4

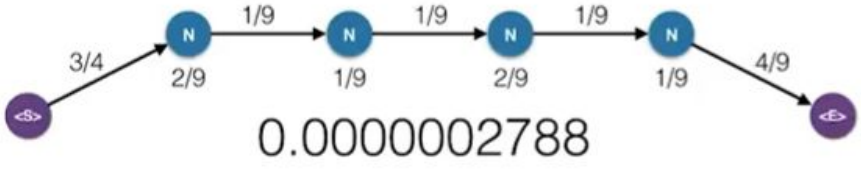




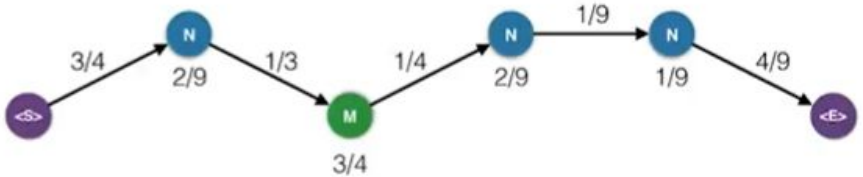




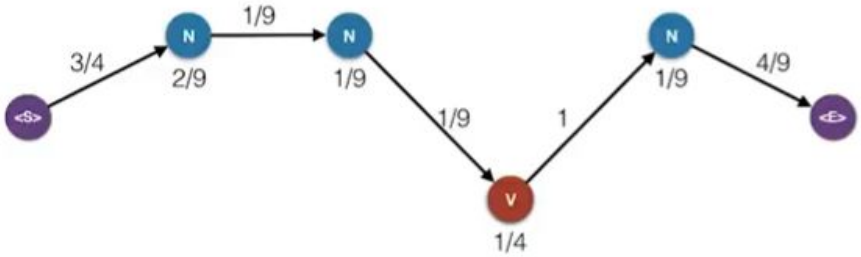
Скрытые марковские модели



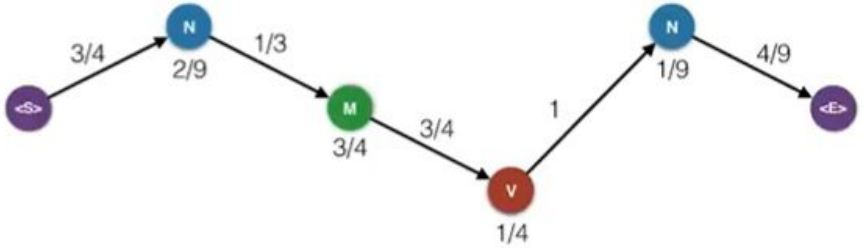
0.0000002788



0.0000127

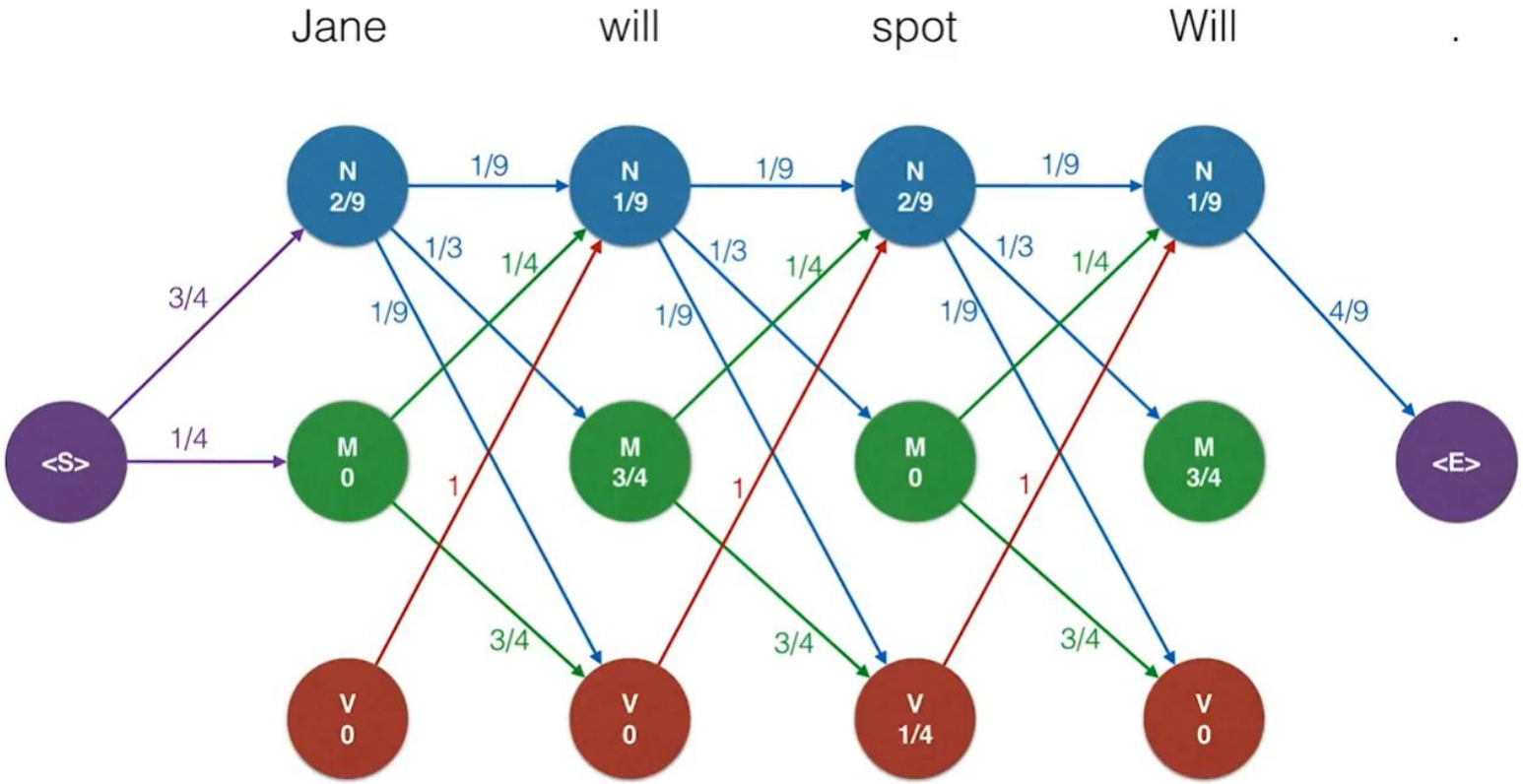


0.00000282



0.000385

Алгоритм Витерби



Алгоритм Витерби

https://www.youtube.com/watch?v=mHEKZ8jv2SY&list=PLC0PzjY99Q_U5bba7gYJicCxIufrFmlTa&index=13

Задача на марковские процессы

В процессе опроса владельцев автомобилей трех американских марок: марки A, марки B, марки C, им был задан вопрос о том, какую торговую марку они бы выбрали для следующей покупки.

Среди владельцев автомобилей марки A 20% сказали что выберут опять эту же марку, 50% сказали, что они бы перешли на марку B%, а 30% заявили, что предпочли бы марку C.

Среди владельцев автомобилей марки B 20% сказали, что перейдут на марку A, в то время как 70% заявили, что приобрели бы опять автомобиль марки B, а 10% заявили, что в следующий раз предпочли бы марку C.

Среди владельцев автомобилей C 30% ответили, что перешли бы на марку A, 30% сказали, что перешли бы на марку B, а 40% заявили, что остались бы верны той же марке C.

Задача на марковские процессы

Вопрос 1 : Если некто приобрел автомобиль марки A, то какова вероятность, что его второй машиной будет автомобиль марки C?

Некоторый fun

Задача:

Даны фразы из биографии французской актрисы Эммануэль Беар, приведённой на сайте «Каталог биографий известных актёров».

Некоторый fun

1. Режиссерам привзглянулась нежная красота Беар, и без ролей она не сидела.

2. Но «своего» режиссера Эммануэль порадостнилось встретить лишь в 1992 году.

3. Обрелась невероятно тонкая и красивая картина (не в последнюю очередность благодаря Беар), которая обрела «Сезара» как оптимальный кинофильм того года.

4. Она нанастолькоко ладно сыграла метания героини между двумя супругчинами, что Даниэль Отёй, который был супругом Беар в кинофильме и в жизни, выбирал не приезжать на съемки, когда там снимались сцены с любовником героини Эммануэль.

Некоторый fun

5. Своих детей и свою личную жизнь артистка ревностно оберегает от внимания газетчиков, но папарацци очень любят Беар, видимо, позжеу что она очень фотогенична.

6. Много лет Эммануэль Беар была «лицом» известной фирмы «Christian Dior», но не так давно ее на этом посту поменяла российская манекенщица Крправда Семеновская.

Задание 1. Отметьте слова, которые вам показались странными.

Задание 2. Объясните их появление в этом тексте.

Некоторый fun

Программки, которые порождают похожие слова, как правило именуют синонимайзерами, и их довольно ценят те, кто наваривает созданием и «раскруткой» вебсайтов. Дело в том, собственно что поисковые системы давным-давно научились отсеивать сайты-близнецы с схожим или же подобным содержанием, и задачка скорого сотворения оригинальных слов на заданную тему довольно популярна. В онлайне просто присутствуют почти все 10-ки этих сервисов. Вот итог обработки сего абзаца одним из них.

Некоторый fun

Некоторый fun

- Подумайте, как NLP помогает отсеивать такие сайты?
- Как вы думаете насколько давно придумали эту задачу?

Задача на языковые модели

Попробуйте описать образование глагольных основ в языке **йоулумни** (индейский язык где-то в Северной Америке). Запишите регулярками 3 основы

основа	герундий	дуратив
caw “кричать”	caw-inay	cawaa-ʔaa-n
cuum “разрушать”	cum-inay	cumuu-ʔaa-n
hoyoo “называть”	hoy-inay	hoyoo-ʔaa-n
diiyl “охранять”	diyl-inay	diyiil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

Глагольные формы в языке йоулумни

Спасибо за внимание!

Литература