

NLP Assignment 3

ياسمن لطف اللهی - ٩٧٥٢١٤٨٦

١.

مرحله اول

Tom	will	Mark	Watch	Will	Can	Mark	Watch	Can	Tom	Watch	Mark	Can	Watch
N	M	V	N	N	M	V	N	M	N	V	N	M	V

Emission probabilities :

Word/Tag	Noun	Verb	Modal
Tom	2/6	0	0
Will	1/6	0	1/4
Mark	1/6	2/4	0
Watch	2/6	2/4	0
Can	0	0	3/4

مرحله دوم

Transition probability

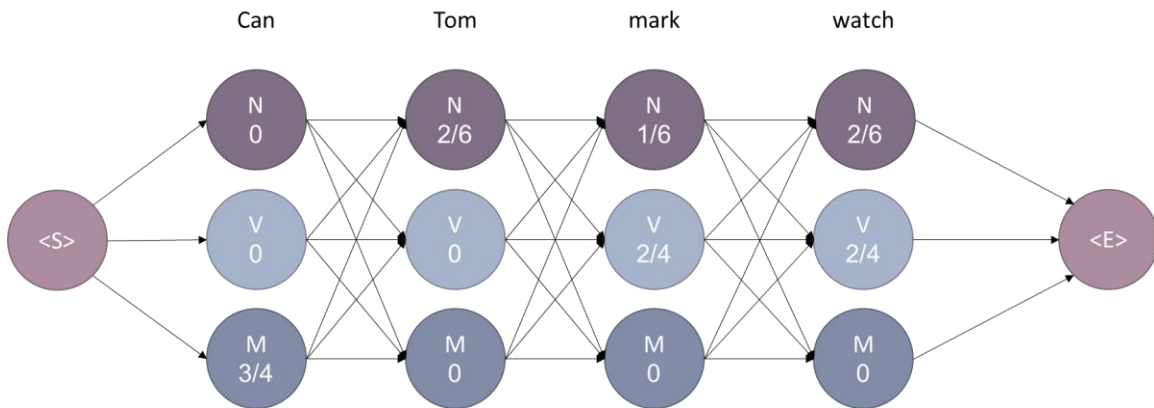
	Noun	Verb	Modal	<E>
<S>	3/4	0	1/4	0
Noun	0	1/6	3/6	2/6
Verb	2/4	0	0	2/4
Modal	1/4	0	3/4	0

مرحله سوم

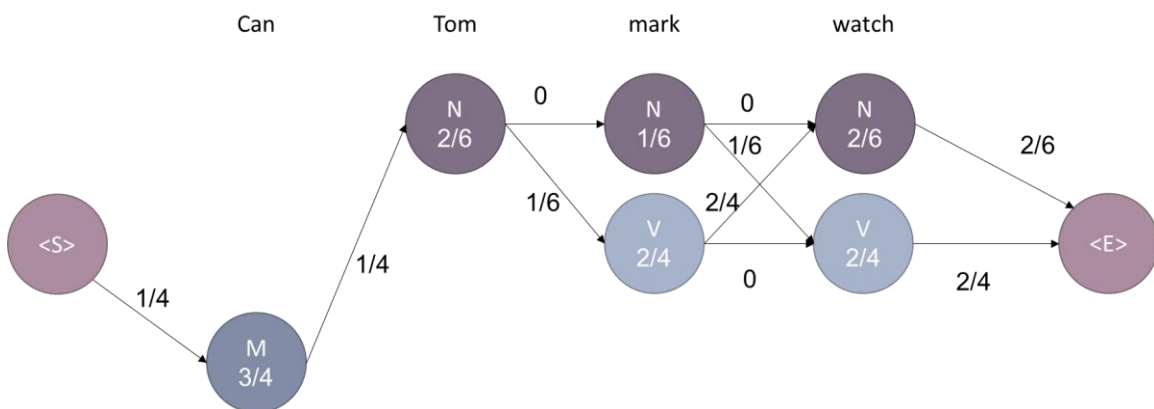
گراف

Can Tom mark watch?

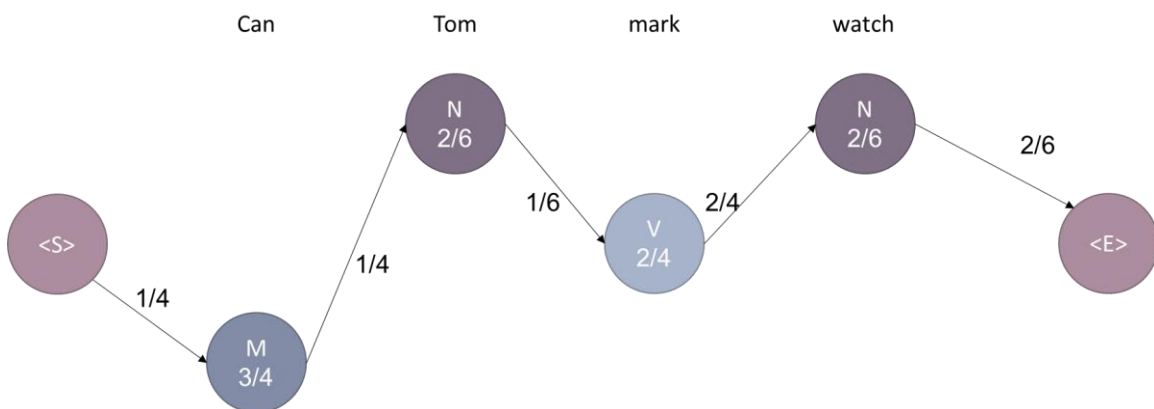
- گراف اولیه



- حذف راس های صفر



- حذف یال های صفر و راس هایی که به <E> نمی رسند



- احتمال جمله

$$P(\text{Can Tom mark watch}) = \frac{1}{4} \times \frac{3}{4} \times \frac{1}{4} \times \frac{2}{6} \times \frac{1}{6} \times \frac{2}{4} \times \frac{2}{4} \times \frac{2}{6} \times \frac{2}{6} = 7.23 \times 10^{-5}$$

0.9241 .۲

John	Eats	Pie	With	Cream	
N (0.2), NP (0.04)	S (0.00384)	S (0.000058)	-	S (0.00000041472)	John
	V (0.3), VP (0.12)	VP (0.018)	-	VP (0.000013)	Eats
		N (0.1), NP (0.02)	-	NP (0.000144)	Pie
			P (0.6)	PP (0.036)	With
				N (0.3), NP (0.06)	Cream

John

- Noun \rightarrow John = 0.2
- NP \rightarrow Noun \rightarrow John = (0.2)(0.2) = 0.04

Eats

- Verb \rightarrow eats = 0.3
- VP \rightarrow Verb \rightarrow eats = (0.4)(0.3) = 0.12

Pie

- Noun \rightarrow Pie = 0.1
- NP \rightarrow Noun \rightarrow Pie = (0.2)(0.1) = 0.02

With

- P \rightarrow with = 0.6

Cream

- $\text{Noun} \rightarrow \text{Cream} = 0.3$
- $\text{NP} \rightarrow \text{Noun} \rightarrow \text{Cream} = (0.2)(0.3) = 0.06$

John eats

- $\text{S} \rightarrow \text{NP VP} = (0.8)(0.04)(0.12) = 0.00384$

Eats pie

- $\text{VP} \rightarrow \text{Verb NP} = (0.3)(0.3)(0.02) = 0.0018$

With cream

- $\text{PP} \rightarrow \text{P NP} = (1)(0.6)(0.06) = 0.036$

John eats pie

- $\text{S} \rightarrow \text{NP VP} = (0.8)(0.04)(0.0018) = 0.000058$

Pie with cream

- $\text{NP} \rightarrow \text{NP PP} = (0.2)(0.02)(0.036) = 0.000144$

Eats pie with cream

- $\text{VP} \rightarrow \text{Verb(eats) NP (pie with cream)} = (0.3)(0.3)(0.000144) = 0.00001296$
- $\text{VP} \rightarrow \text{VP(eats pie) PP (with cream)} = (0.2)(0.018)(0.036) = 0.00001296$

John eats pie with cream

- $\text{S} \rightarrow \text{NP VP} = (0.8)(0.04)(0.000013) = 0.00000041472$

۳.

محدودیت های PCFG

۱. احتمال قوانین همیشه ثابت در نظر گرفته می شود. با اینکه با توجه به کلمات بکار رفته و اطلاعات Lexical، ممکن است احتمال متفاوت باشد. در Lexicalized PCFG احتمال قانون به کلمات نیز

بستگی دارد. برای مثال، در قانون $VP \rightarrow Verb\ NP\ NP$ ، احتمال به اینکه چه فعلی در V استفاده شده باشد، بستگی دارد. مثلاً، احتمال در صورت استفاده از $give$ متفاوت از run است.

۲. از دیگر معایب ثابت بودن احتمال قوانین، استفاده از قوانین در مکان های مختلف است. ممکن است استفاده از یک قانون در ابتدای جمله با انتهای جمله احتمال متفاوتی داشته باشد. در $Lexicalized$ نیز این محدودیت وجود دارد، مگر اینکه علاوه بر اطلاعات $Lexical$ خود قانون و کلماتش، اطلاعات قوانین قبل و بعد را نیز در نظر بگیریم.

سوال عملی

دقت رده بند با آموزش روی ۵۰ هزار داده، ۹۲۴۱/۰ است.

با توجه به ایمپورت های نوت بوک، برای ساخت رده بند، از $Pipeline$ ، $DictVectorizer$ و درخت تصمیم استفاده شد. یک تابع برای تهیه فیچرها از کلمات برای درخت نوشته شد. از تابع $train_test_split$ کتابخانه $sklearn$ برای جداسازی داده آموزش و تست استفاده شد.

در آخر تابع pos_tag نوشته شد تا تگ هارا پیش بینی کند و به صورت خواسته شده برگرداند.