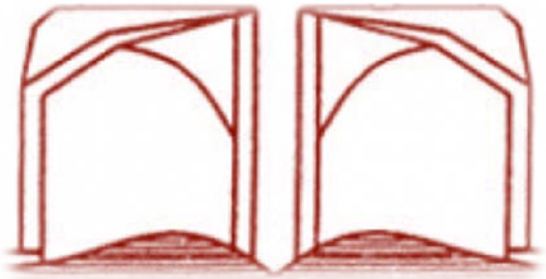


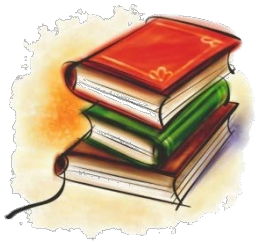
آشنایی با زبان‌شناسی رایانشی

برچسب‌زنی اجزای کلام
(Part-of-Speech Tagging)

هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران - دانشکده علوم و فنون نوین



فهرست

اجزای کلام: Part-of-Speech (POS) . . .

- بیانگر مقوله نحوی که هر کلمه به آن تعلق دارد
- نام‌های مختلف
  - part-of-speech(POS), tags ،lexical tags ،word classes
- مثال
  - {من، تو، او، ...} یک {کتاب، گوسفند، درخت، ...} را {دیدم، خریدم، فروختم، ...}
- برچسب‌زنی (Tagging) = POS Tagging
  - فرایند انتساب مقوله نحوی به هر کلمه در پیکره متنی
- ایده قدیمی است
  - پیشنهاد ۸ دسته توسط Thrax در ۱۰۰ ق.م.
    - noun, verb, pronoun, preposition, adverb, conjunction, participle, article
  - مشابه برچسب‌های مورد استفاده در کتب درسی امروزی
    - noun, verb, adjective, preposition, adverb, conjunction, pronoun, interjection

اجزای کلام: Part-of-Speech (POS) ...

- POSهای اصلی در انگلیسی
  - اسم (Noun)
    - Ali, IBM, Book, ...
  - فعل (Verb)
    - Eat, Make, Draw, ...
  - صفت (Adjective)
    - Good, Beautiful, Young, ...
  - قید (Adverb)
    - Slowly, Very, Fortunately, ...
  - حرف اضافه (Preposition)
    - on, to, by, from, with, ...
  - حرف تعریف (Determiner)
    - a, an, the
  - ضمیر (Pronoun)
    - she, who, my, others, ...
  - حرف ربط (Conjunctions)
    - and, but, if, ...
  - افعال کمکی (Auxiliaries)
    - (can, should, are, ...)
  - حروف (Particles)
    - up, down, off, in, out, ...
  - شماره‌ها(Numerals)
    - one, two, three, first, second, ...

اجزای کلام: Part-of-Speech (POS) . . .

- بسته به کاربرد، ممکن است POSهای جزئی‌تری نیز به کار روند
  - اسم (Noun)
    - فرد یا جمع (Singular و Plural)
  - صفت (Adjective)
    - تفضیلی یا عالی (Comparative یا Superlative)
  - فعل (Verb)
    - اصلی یا وجهی (Main یا Modal)
  - ضمیر (Pronoun)
    - شخصی یا مالکیت (Personal یا Possessive)
  - ...

- **Noun (person, place or thing)**
  - Singular (NN): dog, fork
  - Plural (NNS): dogs, forks
  - Proper (NNP, NNPS): John, Springfields
  - Personal pronoun (PRP): I, you, he, she, it
  - Wh-pronoun (WP): who, what
- **Verb (actions and processes)**
  - Base, infinitive (VB): eat
  - Past tense (VBD): ate
  - Gerund (VBG): eating
  - Past participle (VBN): eaten
  - Non 3rd person singular present tense (VBP): eat
  - 3rd person singular present tense: (VBZ): eats
  - Modal (MD): should, can
  - To (TO): to (to eat)

- برای علائم نقطه‌گذاری هم ممکن است برچسب‌های متفاوتی به کار رود

اجزای کلام: Part-of-Speech (POS) . . .

انواع اجزای کلام (POS: Part-Of-Speech)

- **واژه‌های محتوایی (Content/Lexical Words) = گروه باز = پذیرش اعضای جدید**
  - حامل معنی و استفاده به صورت مستقل
  - در برخی دسته‌ها، اعضای نامتناهی دارند (مانند اسم): تولید اعضای جدید (اسم خاص)

برچسب	توصیف	نقش	مثال
N	(اسم)Noun	نامیدن موجودات	cat
V	(فعل)Verb	نامیدن رویداد یا شرایط	forget
Adj	(صفت)Adjective	توصیفی	yellow
Adv	(قید)Adverb	حالت عملکرد	quickly
Interj	(حرف ندا)Interjection	واکنش	oh!

- **واژه‌های نقشی (Function Words): گروه بسته = اعضای محدود و ثابت**
  - کوتاه، پرتکرار و دارای ساختار نحوی مشخص و استفاده برای بیان رابطه گرامری با سایر کلمات

برچسب	توصیف	نقش	مثال
Conj	(حرف ربط) Conjunction	پیوند دهنده‌ی عبارات	and
Det	(حرف تعریف) Determiner	مشخص‌کننده‌ی شناخته بودن	the
Prep	(حرف اضافه) Preposition	روابط زمانی، مکانی، جهتی	from
Pron	(ضمیر) Pronoun	ارجاع ساده شده	she

ac.ir)

اجزای کلام: Part-of-Speech (POS) . . .

Open class (lexical) words

- Nouns
  - Proper: *IBM, Italy*
  - Common: *cat / cats, snow*
- Verbs
  - Main: *see, registered*
  - Modals: *can, had*
- Adjectives *old older oldest*
- Adverbs *slowly*
- Numbers *122,312, one*
- … more

Closed class (functior

- Determiners *the some*
- Conjunctions *and or*
- Pronouns *he its*
- Prepositions *to with*
- Particles *off up*
- Interjections *Ow Eh*
- … more

اجزای کلام: Part-of-Speech (POS) . . .

Penn Treebank

برچسب	توصیف	مثال
CC	حرف ربط پیوند دهنده	and
CD	عدد شمارشی	two
DT	حرف تعریف	the
EX	عبارت وجود داشتن there	there
FW	واژه‌ی خارجی	omerta
IN	حرف اضافه، حرف ربط پیرو	over, but
JJ	صفت	yellow
JJR	صفت، مقایسه‌ای	better
JJS	صفت، تفضیلی	best
LS	نشانه‌ی عنصر لیست	
MD	فعل وجهی	might
NN	اسم، مفرد یا حجمی	rock, water
NNS	اسم، جمع	rocks
NNP	اسم خاص، مفرد	Joe
NNPS	اسم خاص، جمع	Red Guards
PDT	حرف تعریف پیشین	all (all the girls)
POS	نشانه‌ی ملکی	's
PRP	ضمیر شخصی	I
PRP$	ضمیر ملکی	Mine
RB	قید	Quickly
RBR	قید، مقایسه‌ای	higher
RBS	قید، تفضیلی	highest
RP	وند فعلی	up (take up.)
TO	To	to
UH	حرف ندا	hey!
VB	بن فعل	choose
VBD	فعل، زمان گذشته	chose
VBG	فعل، استمراری یا صفت فاعلی	choosing
VBN	فعل، صفت مفعولی	chosen
VBP	فعل، مفرد غیر سوم شخص زمان حال	jump
VBZ	فعل، سوم شخص مفرد زمان حال	jumps
WDT	حرف تعریف Wh	which
WP	ضمیر Wh	who
WP$	ضمیر ملکی Wh	whose
WRB	قید Wh	when

- مجموعه برچسب‌های (tag set) انگلیسی
  - برچسب‌های Penn Treebank شامل ۴۵ برچسب
    - امکان کاهش تا ۱۲ برچسب رایج
  - برچسب‌های Brown Corpus شامل ۸۷ برچسب
    - پیکره یک میلیون کلمه‌ای از ۵۰۰ نویسنده در ژانرهای مختلف
  - برچسب‌های C5 شامل ۶۱ برچسب
    - در Lancaster UCREL project's CLAWS
  - برچسب‌های C7 شامل ۱۴۶ برچسب
- در پیکره متنی زبان فارسی (دکتر بیجن خان)
  - حدود ۶۶۰ برچسب
    - کاهش تعداد برچسب‌ها به حدود ۴۰ مورد

اجزای کلام: Part-of-Speech (POS) . . .

- برچسب‌های Penn Treebank (رایج در انگلیسی)

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb gerund	*eating*
JJ	adjective	*yellow*	VBN	verb past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, sing.	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

اجزای کلام: Part-of-Speech (POS) . . .

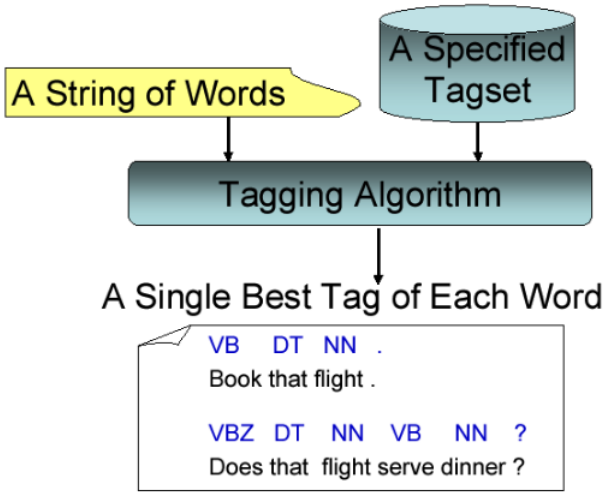
- برچسب‌های پیکره بیجن خان (فارسی)

No	Tag	Description
1	ADJ	Adjective, General
2	ADJ_CMPR	Adjective, Comparative
3	ADJ_INO	Past Participle
4	ADJ_ORD	Adjective, Ordinal
5	ADJ_SIM	Adjective, Simple
6	ADJ_SUP	Adjective, Superlative
7	ADV	Adverb, General
8	ADV_EXM	Adverb, Exemplar
9	ADV_I	Adverb, Question
10	ADV_NEGG	Adverb, Negation
11	ADV_NI	Adverb, Not Question
12	ADV_TIME	Adverb, Time
13	AR	Arabic Word
14	CON	Conjunction
15	DEFAULT	Default
16	DELM	Delimiter
17	DET	Determiner
18	IF	Conditional
19	INT	Interjection
20	MORP	Morpheme

No	Tag	Description
21	MQUA	Modifier of Quantifier
22	MS	Mathematic Symbol
23	N_PL	Noun, Plural
24	N_SING	Noun, Singular
25	NN	Number
26	NP	Noun Phrase
27	OH	Oh Interjection (حرف ندا)
28	OHH	Oh noun (منادی)
29	P	Preposition
30	PP	Prepositional Phrase
31	PRO	Pronoun
32	PS	Psedo-Sentence
33	QUA	Quantifier
34	SPEC	Specifier
35	V_AUX	Verb, Auxiliary
36	V_IMP	Verb, Imperative
37	V_PA	Verb, Past Tense
38	V_PRE	Verb, Predicative
39	V_PRS	Verb, Present Tense
40	V_SUB	Verb, Subjunctive

برچسب‌زنی اجزای کلام (POS Tagging)

- یافتن برچسب کلمات در متن



Penn Treebank POS tags

- Input: Plays well with others
- Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS
- Output: Plays/VBZ well/RB with/IN others/NNS

- ابتدا باید متن را نرمال کرد و علائم نگارشی را جدا تفکیک کرد

برچسب‌زنی اجزای کلام: کاربردها

- مدل‌سازی زبانی (در بازشناسی گفتار و ...)
  - استفاده در پیش بینی کلمه بعدی
    - مثال: در انگلیسی بعد از ضمایر ملکی، اسم و بعد از ضمایر شخصی، فعل می‌آید.
    - در فارسی معمولاً بعد از حرف اضافه اسم یا ضمیر می‌آید.
- سنتز گفتار: حاوی اطلاعاتی در مورد نحوه تلفظ صحیح یک کلمه
  - مثال: تلفظ کلمه object در انگلیسی به صورت OBject (noun) و obJECT (verb)
  - تلفظ کلمه "مرد" در فارسی به صورت "مَرد" (اسم)، "مُرد" (فعل)
- بازیابی اطلاعات: کمک به استخراج کلمات مهم در متن
- رفع ابهام معنایی
  - کمک به رفع ابهام معنایی (استفاده در سیستم‌های ترجمه و ...)
    - مثال: کلمه watch در انگلیسی و کلمه "در" در فارسی
- تجزیه نحوی(parsing)
  - در تجزیه پایین به بالا، نیاز به دانستن مقوله نحوی کلمات است

برچسب‌زنی اجزای کلام: چالش . . .

- بعضی از کلمات به چند مقوله نحوی تعلق دارند
  - حدود ۴۰٪ از کلمات (Token) پیکره Brown دارای بیش از یک برچسب هستند
    - حدود ۱۱.۵٪ از انواع کلمات (Word Type) به کار رفته در پیکره را تشکیل می‌دهند

		Original 87-tag corpus		Treebank 45-tag corpus	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*'s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

ابهام انواع کلمات در پیکره Brown

برچسب‌زنی اجزای کلام: چالش . . .

- مثال از ابهام در برچسب زدن

- Mrs/NNP Shaefer/NNP never/RB got/VBD **around/RP** to/TO joining/VBG
- All/DT we/PRP gola/VBN do/VB is/VBZ go/VB **around/IN** the/DT corner/NN
- Chateau/NNP Petrus/NNP costs/VBZ **around/RB** 250/CD

- I know **that** he is honest = IN (relativizer)
- Yes, **that** play was nice = DT (determiner)
- You can't go **that** far = RB (adverb)

برچسب‌زنی اجزای کلام: منابع اطلاعات (ویژگی‌ها) . . .

- از چه اطلاعاتی برای برچسب‌زنی اجزای کلام می‌توان استفاده کرد؟
  - خود کلمه (به تنهایی)
    - Word — the: the → DT
    - حرف بزرگ در اول کلمه (انگلیسی)
      - کلمات با حرف اول بزرگ، معمولا به عنوان اسم (خاص) برچسب زده می‌شوند اما برای کلمات اول جمله بهتر است حالت حروف کوچک آن هم بررسی شود
        - Lowercased word — Importantly: importantly → RB
    - وجود برخی پسوندها و پیشوندها: بیانگر نوع کلمه است
      - Prefixes — unfathomable: un- → JJ
      - Suffixes — Importantly: -ly → RB
    - حالت حروف بزرگ کلمات
      - Capitalization — Meridian: CAP → NNP
    - ساختار کلمه (عبارت)
      - Word shapes — 35-year: d-x → JJ

برچسب‌زنی اجزای کلام: منابع اطلاعات (ویژگی‌ها)

- از چه اطلاعاتی برای برچسب‌زنی اجزای کلام می‌توان استفاده کرد؟
  - کلمات همسایه و دنباله کلمات
    - Bill saw that man yesterday
    - NNP NN DT NN NN
    - VB VB(D) IN VB NN
    - در این مثال، ترکیب DT VB (حرف تعریف قبل از فعل) کمتر رخ می‌دهد
  - احتمال برچسب‌ها/کلمات
    - احتمال اینکه کلمه man فعل (VB) باشد، خیلی کم است
    - احتمال پست سر هم آمدن برچسب‌های مختلف پشت سر هم
      - در فارسی: احتمال اینکه بعد از «اسم» برچسب «صفت» بیاید زیاد است اما عکس آن کم است

برچسب‌زنی اجزای کلام: روش‌ها

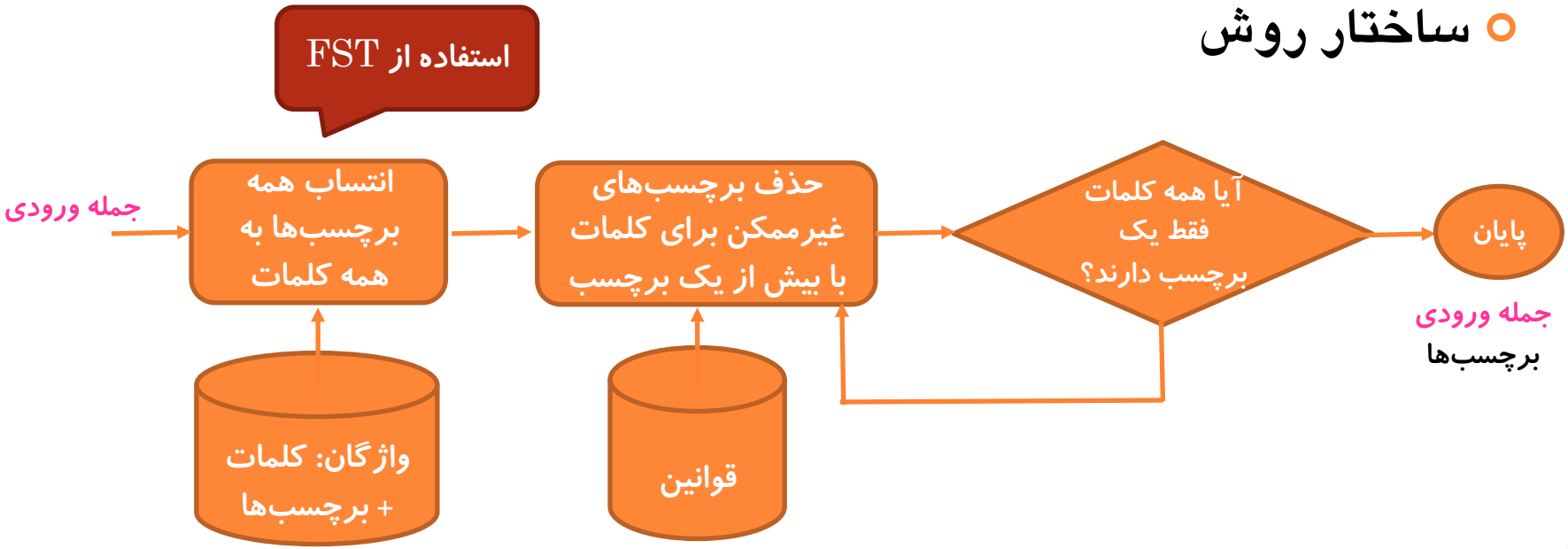
- **روش‌های مبتنی بر قاعده (Rule-Based POS Tagging)**
  - **استفاده از یک واژگان با برچسب(های) اجزای کلام برای هر کلمه و قوانینی برای رفع ابهام**
  - **مثال:** EngCG (Constraint Grammar)
- **روش‌های مبتنی بر یادگیری از داده**
  - **آماری** (Probabilistic/Stochastic POS Tagging)
    - استفاده از احتمال برای انتخاب برچسب‌ها
    - استفاده از روش‌های یادگیری آماری و یک پیکره دارای برچسب برای محاسبه احتمال‌ها
    - مثال: HMM Tager
  - **شبکه عصبی مصنوعی** (Artificial Neural Network)
- **ترکیبی: مبتنی بر تبدیل (Transformation-Based POS Tagging)**
  - **استفاده از قواعد برای رفع ابهام (روش‌های مبتنی بر قاعده)**
  - **استفاده از یادگیری ماشین برای استخراج خودکار قوانین از پیکره متنی**
  - Brill's tagger

برچسب‌زنی اجزای کلام: روش مبتنی بر قاعده . . .

نیازمندی‌ها

- واژگانی حاوی کلمات و همه برچسب‌های ممکن برای هر کلمه
- مجموعه قوانینی برای حذف برچسب‌های اضافی انتساب داده شده به یک کلمه
  - حذف برچسب‌های غیرممکن تا فقط یک برچسب برای هر کلمه باقی بماند

ساختار روش



برچسب‌زنی اجزای کلام: روش مبتنی بر قاعده . . .

- **مثال . . .**
  - **جمله ورودی:** She promised to back the bill
  - **واژگان با برچسب‌ها**
    - she: PRP
    - promised: VBN,VBD
    - to: TO
    - back: VB, JJ, RB, NN
    - the: DT
    - bill: NN, VB
  - **گام ۱: انتساب همه برچسب‌های ممکن به کلمات**

	VBN		NN RB JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

برچسب‌زنی اجزای کلام: روش مبتنی بر قاعده . . .

- مثال
  - گام ۲: اعمال قوانین برای حذف برخی از برچسب‌ها

E.g., *Eliminate VBN if VBD is an option when VBN|VBD follows “<start> PRP”*

			NN RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

E.g., *Eliminate VBN if VBD is an option when VBN|VBD follows “<start> PRP”*

			NN RB		
			JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

برچسب‌زنی اجزای کلام: روش مبتنی بر قاعده

Word	POS	Additional POS features
smaller	ADJ	COMPARATIVE
entire	ADJ	ABSOLUTE ATTRIBUTIVE
fast	ADV	SUPERLATIVE
that	DET	CENTRAL DEMONSTRATIVE SG
all	DET	PREDETERMINER SG/PL QUANTIFIER
dog's	N	GENITIVE SG
furniture	N	NOMINATIVE SG NOINDEFDETERMINER
one-third	NUM	SG
she	PRON	PERSONAL FEMININE NOMINATIVE SG3
show	V	PRESENT -SG3 VFIN
show	N	NOMINATIVE SG
shown	PCP2	SVOO SVO SV
occurred	PCP2	SV
occurred	V	PAST VFIN SV

- روش EngCG ENGTWOL
  - استفاده از یک واژگان با ۵۶۰۰۰ ریشه

- به کارگیری یک Lexicon FST برای انتساب برچسب‌ها به کلمات
  - برای جمله . . . Pavlov had shown that salivation

Pavlov	**PAVLOV N NOM SG PROPER**
had	**HAVE V PAST VFIN SVO**
	HAVE PCP2 SVO
shown	**SHOW PCP2 SVOO SVO SV**
that	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation	**N NOM SG**
...	

- استفاده از تعدادی زیادی محدودیت (۳۷۴۴ قانون) برای کاهش برچسب‌ها

ADVERBIAL-THAT RULE
Given input: "that"
if
(+1 A/ADV/QUANT); */* if next word is adj, adverb, or quantifier */*
(+2 SENT-LIM); */* and following which is a sentence boundary, */*
(NOT -1 SVOC/A); */* and the previous word is not a verb like */*
/ 'consider' which allows adjs as object complements */*
then eliminate non-ADV tags
else eliminate ADV tag

برچسب‌زنی اجزای کلام: روش آماری . . .

- ایده
  - در نظر گرفتن احتمال وقوع برچسب‌ها (برچسب‌های محتمل) برای کلمات
  - با فرض داشتن دنباله کلمات $W=w_1 \ldots w_n$، دنباله برچسب‌های $T=t_1 \ldots t_n$ را طوری پیدا کنید که $P(T \mid W)$ بیشینه شود

$$\hat{T} = \arg\max_T P(T \mid W)$$

قانون بیز

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

حذف $P(W)$ تغییری در بیشینه کردن ایجاد نمی‌کند

$$\hat{T} = \arg\max_T P(T \mid W) = \arg\max_T \frac{P(W \mid T)P(T)}{P(W)}$$

$$= \arg\max_T P(W \mid T)P(T) = \arg\max_T P(w_1 w_2 \cdots w_n \mid t_1 t_2 \cdots t_n) P(t_1 t_2 \cdots t_n)$$

Likelihood Prior

- نحوه محاسبه؟

برچسب‌زنی اجزای کلام: روش آماری . . .

- برای محاسبه نیاز به فرض‌های ساده کننده است

$$\hat{T} = \arg\max_{T} P(w_1 w_2 \cdots w_n \mid t_1 t_2 \cdots t_n) P(t_1 t_2 \cdots t_n)$$

- فرض اول: احتمال وقوع یک کلمه فقط به برچسب آن کلمه وابسته است و مستقل از سایر کلمات و برچسب‌های اطراف آن است

$$P(w_1 w_2 \cdots w_n \mid t_1 t_2 \cdots t_n) \approx \prod_{i=1}^{n} P(w_i \mid t_i)$$

- فرض دوم: دیدن یک برچسب فقط به برچسب قبلی آن وابسته است (Bi-Gram)

$$P(t_1 t_2 \cdots t_n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

- بنابر این داریم:

$$P(w_1 w_2 \cdots w_n \mid t_1 t_2 \cdots t_n) P(t_1 t_2 \cdots t_n) \approx \prod_{i=1}^{n} P(w_i \mid t_i) \cdot \prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

برچسب‌زنی اجزای کلام: روش آماری . . .

- مفهوم احتمال $P(t_i \mid t_{i-1})$ (Tag Transition Probability)
  - احتمال آمدن یک برچسب (t_i) بعد از برچسب دیگر (t_{i-1})
  - احتمال آمدن «اسم» (NN) یا «صفت» (JJ) بعد از «حرف تعریف» (DT) بالاست
    - The beautiful story
  - در پیکره Brown داریم: $P(NN \mid DT)=0.49$

- مفهوم احتمال $P(w_i \mid t_i)$ (Word Likelihood)
  - اگر دنبال کلمه‌ای با برچسب t_i هستیم، احتمال اینکه آن کلمه w_i باشد
  - $P(is \mid VBZ)$= احتمال اینکه کلمه با برچسب «VBZ» (فعل حال سوم شخص مفرد)، کلمه «is» باشد
    - در پیکره Brown داریم: $P(is \mid VBZ)=0.47$

برچسب‌زنی اجزای کلام: روش آماری . . .

محاسبه احتمال‌ها

- نیاز به یک پیکره متنی داریم که در آن کلمات دارای برچسب باشند
- محاسبه احتمال $P(t_i \mid t_{i-1})$ (Tag Transition Probability)

$$P(t_i \mid t_{i-1}) = \frac{C(t_{i-1}t_i)}{C(t_{i-1})}$$

- مثال: برای محاسبه $P(NN \mid DT)$ (در پیکره Brown)- برچسب DT به تعداد 116,454 آمده است که بعد از 56,509 مورد از آنها NN آمده است. پس

$$P(NN \mid DT) = \frac{56509}{116454} = 0.49$$

- محاسبه احتمال $P(w_i \mid t_i)$ (Word Likelihood)

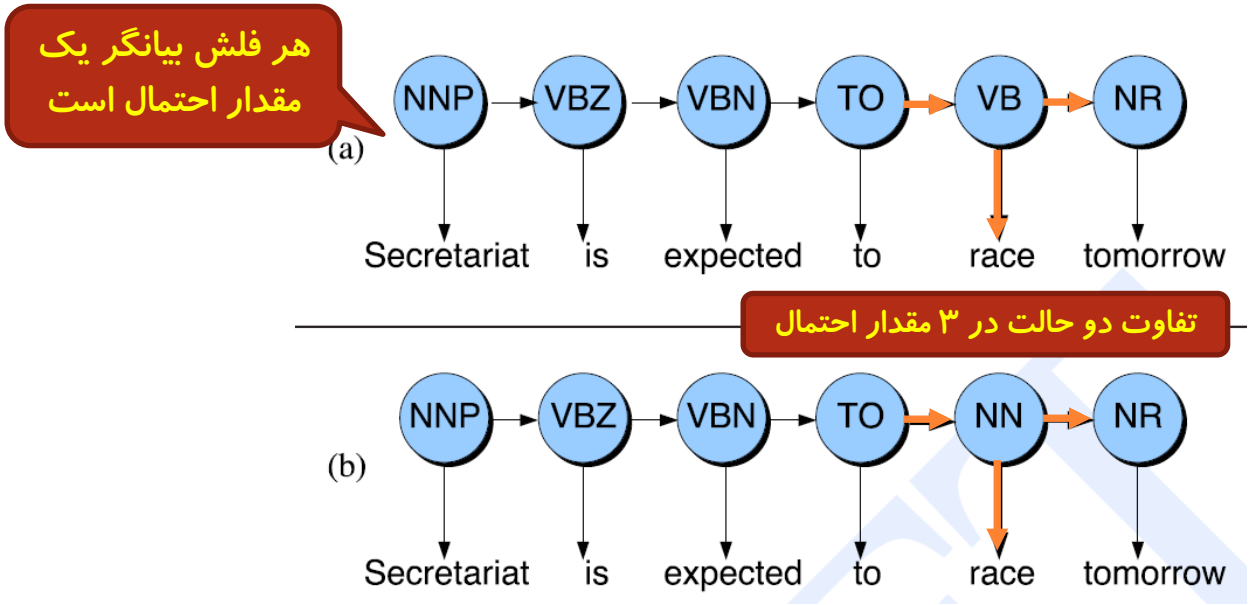
$$P(w_i \mid t_i) = \frac{C(w_i, t_i)}{C(t_i)}$$

- مثال: محاسبه $P(is \mid VBZ)$ (در پیکره Brown)- برچسب «VBZ» به تعداد 21,627 بار آمده که از میان آنها، تعداد 10,073 کلمه «is» است. پس

$$P(is \mid VBZ) = \frac{10073}{21627} = 0.47$$

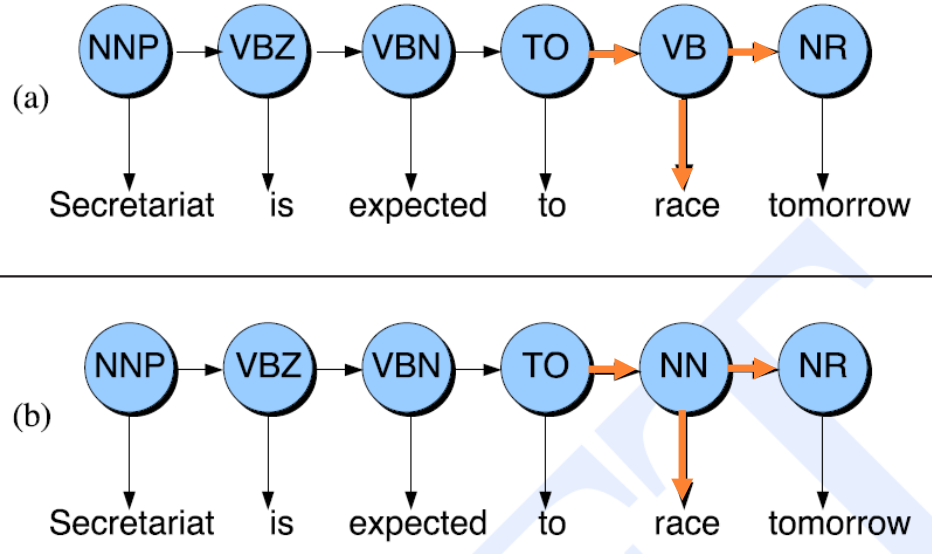
برچسب‌زنی اجزای کلام: روش آماری . . .

- مثال: محاسبه احتمال دنباله برای تعیین برچسب درست
  - کلمه race هم می‌تواند فعل (VB) باشد و هم اسم (NN)
    - Secretariat/NNP is/BEZ expected/VBN to/TO race/VB tomorrow/NR
  - در نظر گرفتن دو حالت از دنباله حالات برای تعیین برچسب race در جمله اول



برچسب‌زنی اجزای کلام: روش آماری . . .

- مثال: محاسبه احتمال دنباله برای تعیین برچسب درست
  - Secretariat/NNP is/BEZ expected/VBN to/TO race/VB tomorrow/NR



$$P(\text{NN}|\text{TO}) = .00047$$
$$P(\text{VB}|\text{TO}) = .83$$

$$P(\text{NR}|\text{VB}) = .0027$$
$$P(\text{NR}|\text{NN}) = .0012$$

$$P(\text{race}|\text{NN}) = .00057$$
$$P(\text{race}|\text{VB}) = .00012$$

مقدار احتمال برای برچسب VB بیشتر است

$$P(\text{VB}|\text{TO})P(\text{NR}|\text{VB})P(\text{race}|\text{VB}) = .00000027$$
$$P(\text{NN}|\text{TO})P(\text{NR}|\text{NN})P(\text{race}|\text{NN}) = .00000000032$$

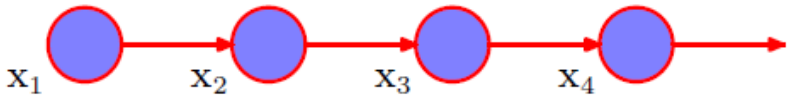
- موارد ابهام دیگر
  - excepted می‌تواند صفت (JJ)، فعل گذشته (VBD) یا اسم مفعول (VBN) باشد

برچسب‌زنی اجزای کلام: روش آماری . . .

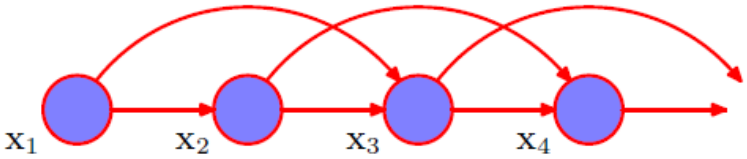
- محاسبه محتمل‌ترین دنباله از برچسب‌ها
  - ساده‌ترین روش: درنظر گرفتن تمام دنباله‌های محتمل و محاسبه احتمال هر یک به روش بیان شده (Brute Force Search)
  - با فرض داشتن N برچسب و T کلمه، حداکثر N^T دنباله از برچسب‌ها تولید می‌شود.
    - محاسبات بسیار زیاد
  - روش‌های رایج
    - مدل مخفی مارکوف (HMM: Hidden Markov Model)
    - میدان تصادفی شرطی (CRF: Conditional Random Field)

برچسب‌زنی اجزای کلام: روش آماری . . .

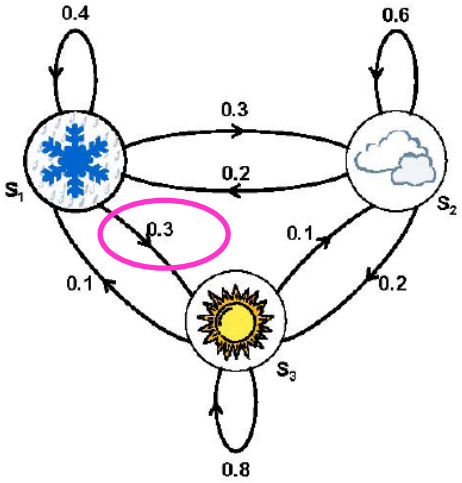
- زنجیره مارکوف (Markov Chain)
  - نوع خاصی از ماشین حالت متناهی (FSA) که در آن به حرکت بین حالت‌ها یک احتمال نسبت داده می‌شود.
  - در نظر گرفتن وابستگی بین حالت‌ها
    - درجه اول (first-order Markov chain): هر حالت تنها به یک حالت قبل وابسته است



    - درجه دوم (second-order Markov chain): هر حالت به دو حالت قبل وابسته است



برچسب‌زنی اجزای کلام: روش آماری . . .



- زنجیره مارکوف: مثال پیش‌بینی وضعیت هوا . . .
  - در نظر گرفتن ۳ حالت (state) مختلف
    - حالت ۱ (S_1): بارندگی (برف یا باران)
    - حالت ۲ (S_2): ابری
    - حالت ۳ (S_3): آفتابی

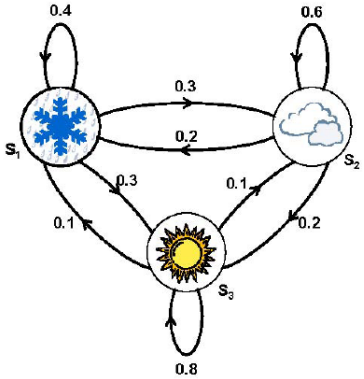
- در نظر گرفتن احتمال انتقال حالت‌ها (state transition probability)

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

احتمال رفتن از حالت ۱ (بارندگی) به حالت ۳ (آفتابی)

- سوال ۱: با فرض اینکه امروز آفتابی است، احتمال اینکه هوای ۷ روز آینده به صورت زیر باشد، چقدر است؟ {آفتابی، آفتابی، باران، باران، آفتابی، ابری، آفتابی}

برچسب‌زنی اجزای کلام: روش آماری . . .



- زنجیره مارکوف: مثال پیش‌بینی وضعیت هوا . . .

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

- سوال ۱: امروز آفتابی است، احتمال اینکه هوای ۷ روز آینده به صورت زیر باشد، چقدر است؟ {آفتابی، آفتابی، باران، باران، آفتابی، ابری، آفتابی}
  - استفاده از وابستگی درجه ۱ (وضعیت هر روز به روز قبل)

مشاهده

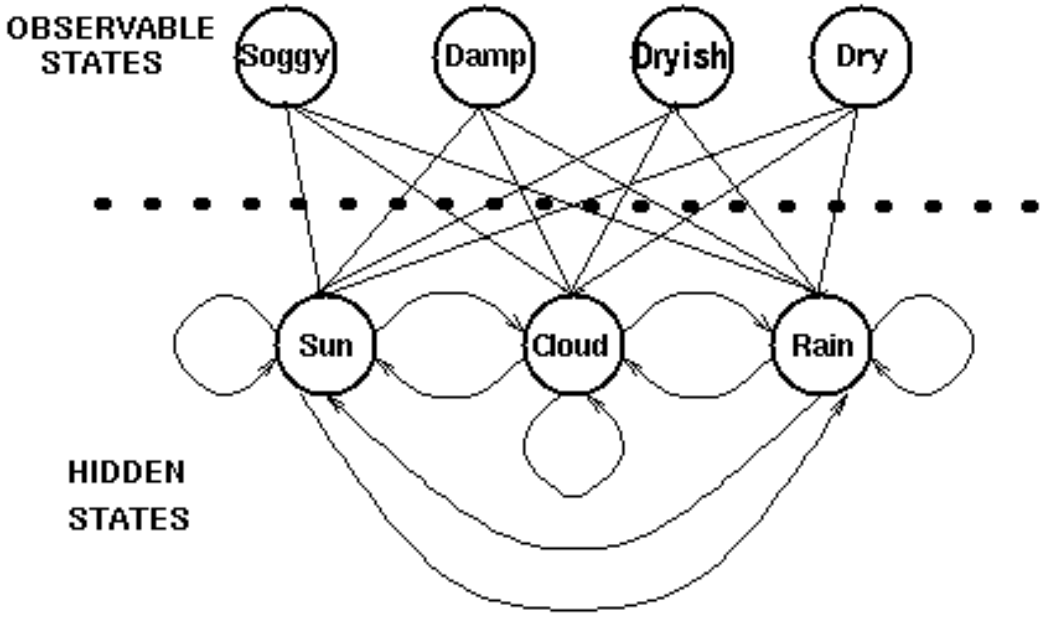
شامل حالت‌ها و ارتباط بین آنها

$$P(O|\text{Model}) = P[S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3|\text{Model}]$$
$$= P[S_3] \cdot P[S_3|S_3] \cdot P[S_3|S_3] \cdot P[S_1|S_3] \cdot P[S_1|S_1] \cdot P[S_3|S_1] \cdot P[S_2|S_3] \cdot P[S_3|S_2]$$
$$= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23}$$
$$= 1 \cdot (0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2)$$
$$= 1.536 \times 10^{-4}$$

احتمال حالت اولیه
(initial state probability)

برچسب‌زنی اجزای کلام: روش آماری . . .

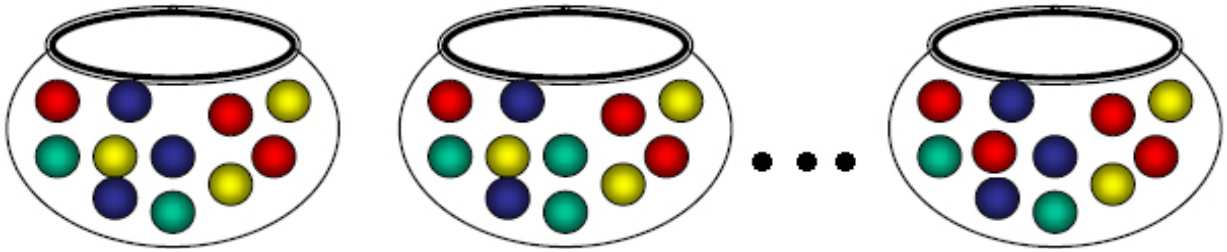
- زنجیره مارکوف: مثال پیش‌بینی وضعیت هوا
  - در آنچه تاکنون بیان شد: حالت‌ها با مشاهده‌ها یکسان بودند
  - در بسیاری از کاربردها، مشاهده‌ها با حالت‌های مساله یکی نیستند
  - حالت‌های اصلی مخفی هستند و باید مشاهده‌ها را با آنها متناظر کرد



مدل مخفی مارکوف
Hidden Markov Model

برچسب‌زنی اجزای کلام: روش آماری . . .

- مدل مخفی مارکوف: مثال گوی و گلدان . . .
  - فرض کنید تعداد N گلدان در یک اتاق داریم
  - در هر گلدان تعداد زیادی گوی رنگی، شامل M رنگ وجود دارد
    - انتخاب هر گوی در در هر گلدان متناسب با مقدار احتمال مرتبط است



URN 1

P(RED) = $b_1(1)$
P(BLUE) = $b_1(2)$
P(GREEN) = $b_1(3)$
P(YELLOW) = $b_1(4)$
⋮
P(ORANGE) = $b_1(M)$

URN 2

P(RED) = $b_2(1)$
P(BLUE) = $b_2(2)$
P(GREEN) = $b_2(3)$
P(YELLOW) = $b_2(4)$
⋮
P(ORANGE) = $b_2(M)$

URN N

P(RED) = $b_N(1)$
P(BLUE) = $b_N(2)$
P(GREEN) = $b_N(3)$
P(YELLOW) = $b_N(4)$
⋮
P(ORANGE) = $b_N(M)$

برچسب‌زنی اجزای کلام: روش آماری . . .

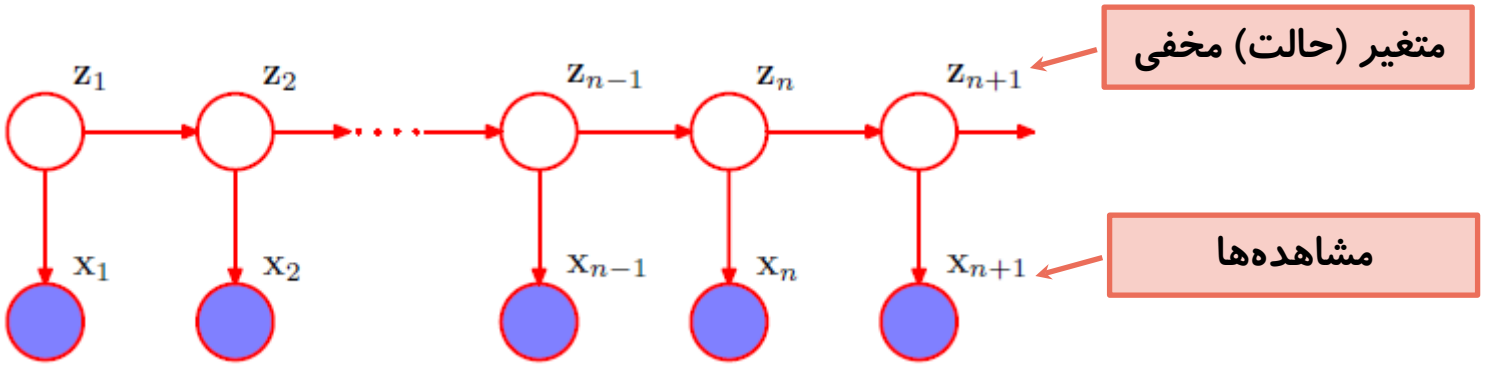
- مدل مخفی مارکوف: مثال گوی و گلدان
  - تعداد N گلدان و M رنگ
  - فرآیند
    - یک نفر (در اتاقی که ما نمی‌بینیم)، یکی از گلدان‌ها را به صورت تصادفی انتخاب می‌کند
    - از داخل گلدان انتخاب شده، یک گوی را بیرون آورده و رنگ آن را اعلام می‌کند
    - گوی را به داخل گلدان مربوطه برمی‌گرداند
    - بر اساس مقداری تصادفی وابسته به گلدان فعلی، گلدان بعدی انتخاب می‌شود
    - مراحل فوق به صورت متوالی تکرار می‌شود
  - دنباله مشاهده: دنباله گوی‌ها (رنگ‌ها)

O = {GREEN, GREEN, BLUE, RED, YELLOW, RED, ..., BLUE}

  - حالت‌ها: گلدان‌ها (از دید مشاهده کننده مخفی است)
  - انتقال حالت‌ها: فرآیند انتخاب گلدان‌ها

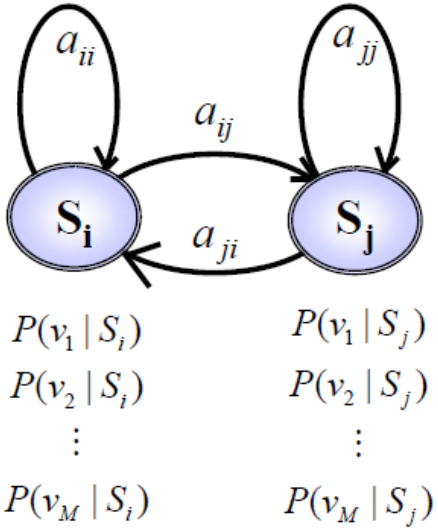
برچسب‌زنی اجزای کلام: روش آماری . . .

- مدل مخفی مارکوف: مبانی
  - مشاهده‌ها توابع احتمالاتی از حالت‌ها هستند
  - دنباله حالت‌ها قابل مشاهده نیستند (مخفی هستند)
  - فرض وابستگی درجه اول
  - مشاهده‌ها فقط به حالت‌ها وابسته هستند و نه به همدیگر



برچسب‌زنی اجزای کلام: روش آماری . . .

- مدل مخفی مارکوف: عناصر اصلی



1. مجموعه‌ای از N حالت
   - در گوی و گلدان: گلدان‌ها

$$S = \{S_1, \cdots S_N\}$$

2. مجموعه‌ای از M نماد مشاهده
   - در گوی و گلدان: رنگ‌ها

$$V = \{v_1, v_2, \cdots v_M\}$$

3. احتمال انتقال حالت‌ها

$$a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$$

   - در گوی و گلدان: جابجایی از یک گلدان به گلدان دیگر

4. احتمال اولیه حالت‌ها

$$\pi = \{\pi_i\} = P(q_1 = i)$$

   - در گوی و گلدان: احتمال انتخاب هر کدام از گلدان‌ها در اولین مشاهده (زمان t=1)

5. تابع توزیع برای نماد kام در حالت jام

$$b_j(k) = P(o_t = v_k \mid q_t = j)$$

   - در گوی و گلدان: احتمال انتخاب گوی kام از گلدان jام
   - تابع توزیع مشاهده‌ها (مثلاً گاوسی)- احتمال تولید مشاهده $o_t=v_k$ در حالت $q_t=j$

- نمایش یک مدل مخفی مارکوف

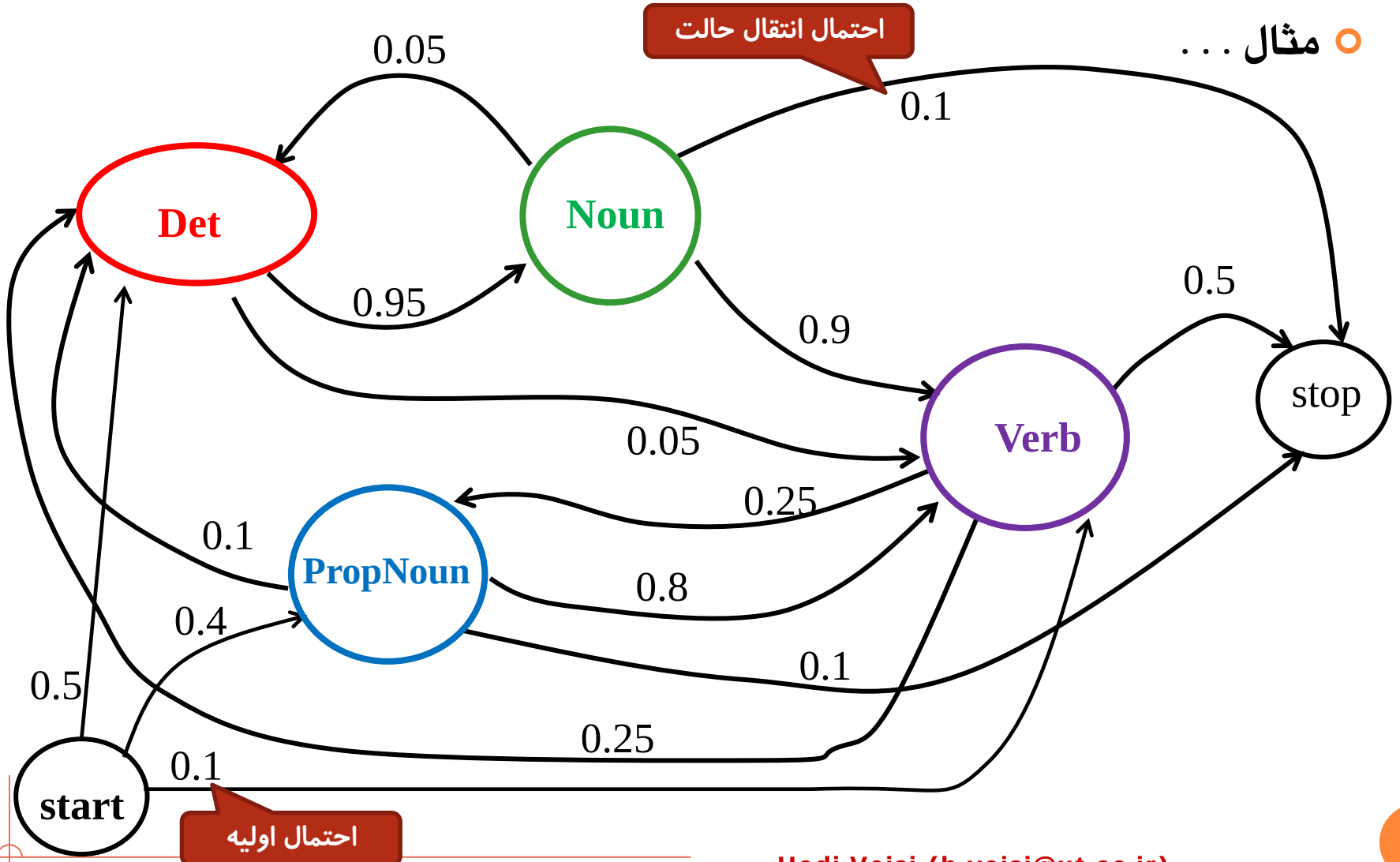
$$\lambda = (A, B, \pi)$$

برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مدل مخفی مارکوف در برچسب‌زنی اجزای کلام
  1. مجموعه‌ای از N حالت = هر حالت بیانگر یک برچسب
     - در گوی و گلدان: گلدان‌ها
  2. مجموعه‌ای از M نماد مشاهده = هر مشاهده بیانگر یک کلمه
     - در گوی و گلدان: رنگ‌ها
  3. احتمال انتقال حالت‌ها = احتمال وقوع یک برچسب بعد از دیگری
     - در گوی و گلدان: جابجایی از یک گلدان به گلدان دیگر
  4. احتمال اولیه حالت‌ها = احتمال اینکه اولین کلمه چه برچسبی داشته باشد
     - در گوی و گلدان: احتمال انتخاب هر کدام از گلدان‌ها در اولین مشاهده (زمان $t=1$)
  5. تابع توزیع برای مشاهده kام در حالت jام = احتمال اینکه برچسب jام کلمه kام باشد
     - در گوی و گلدان: احتمال انتخاب گوی kام از گلدان jام
     - تابع توزیع مشاهده‌ها (مثلاً گاوسی)- احتمال تولید مشاهده $o_t=v_k$ در حالت $q_t=j$

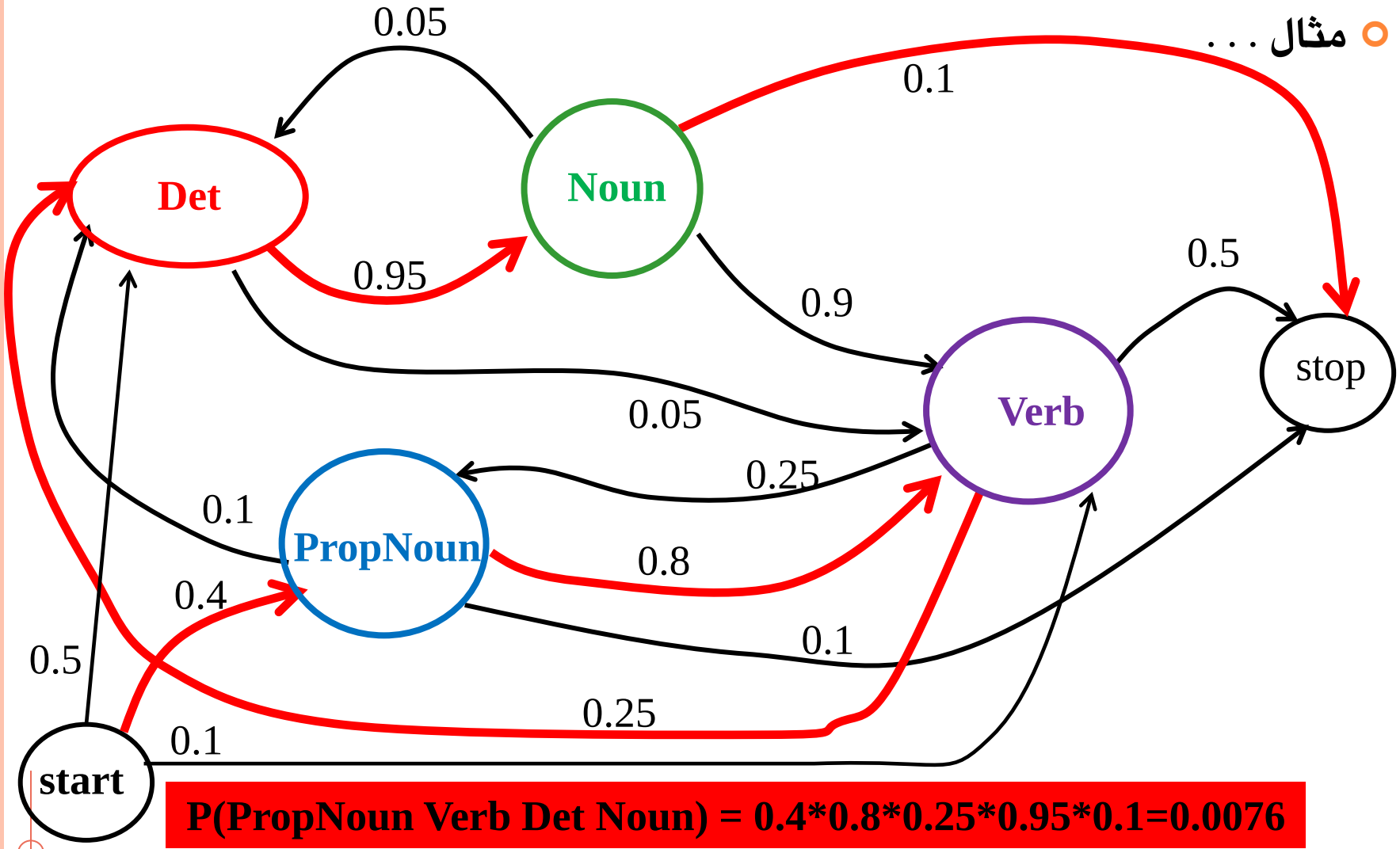
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال . . .



برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

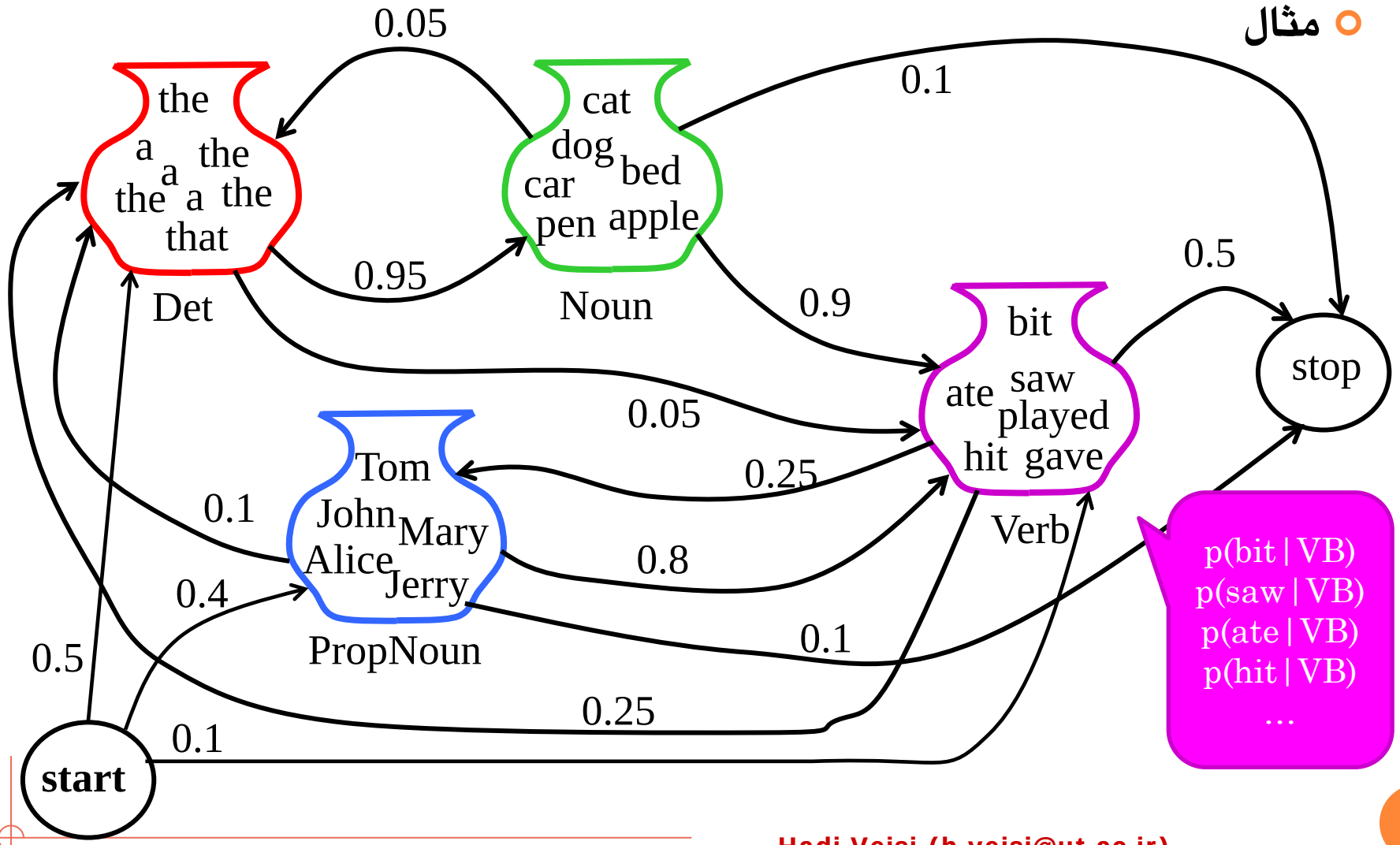
- مثال . . .



P(PropNoun Verb Det Noun) = 0.4*0.8*0.25*0.95*0.1=0.0076

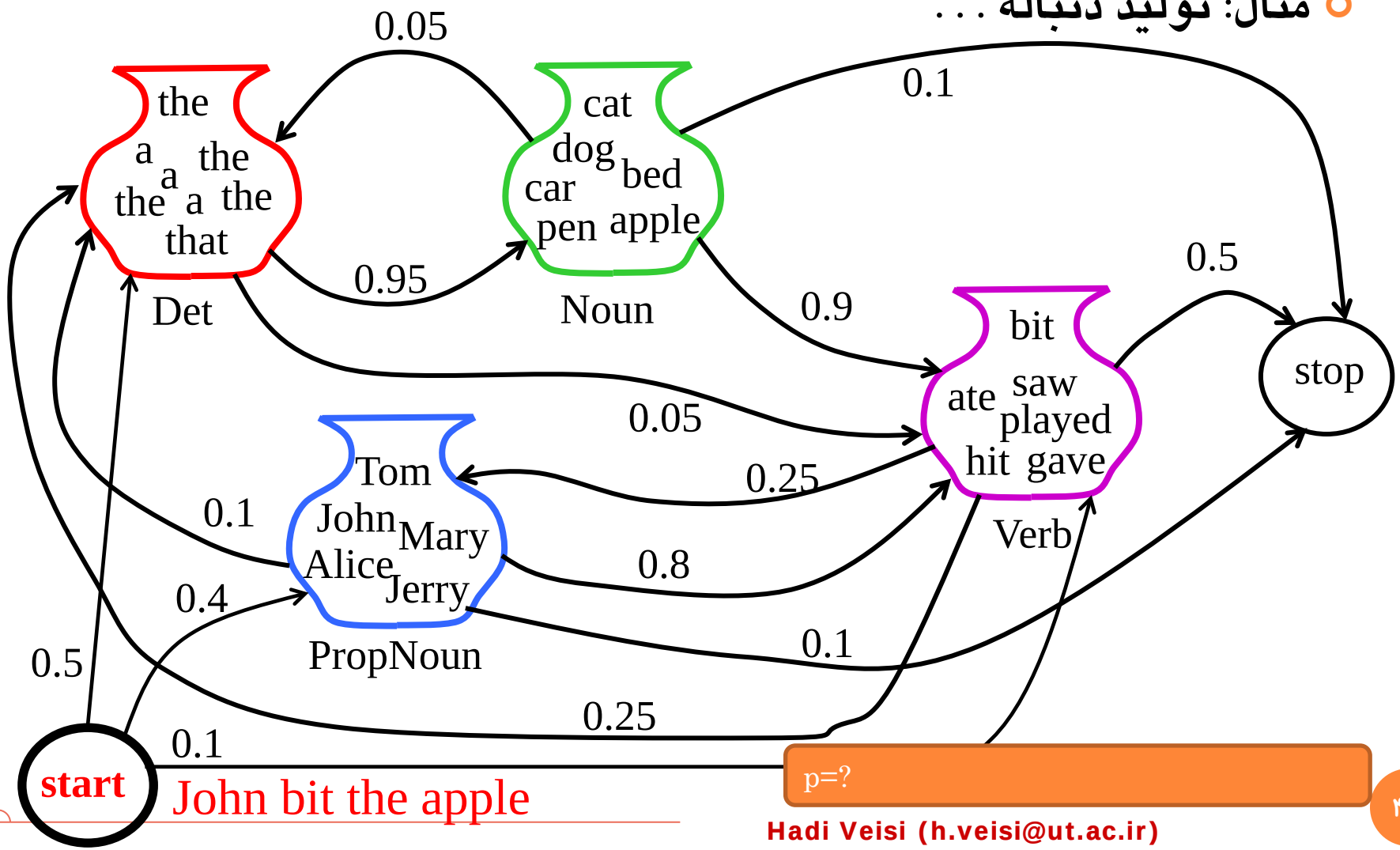
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال



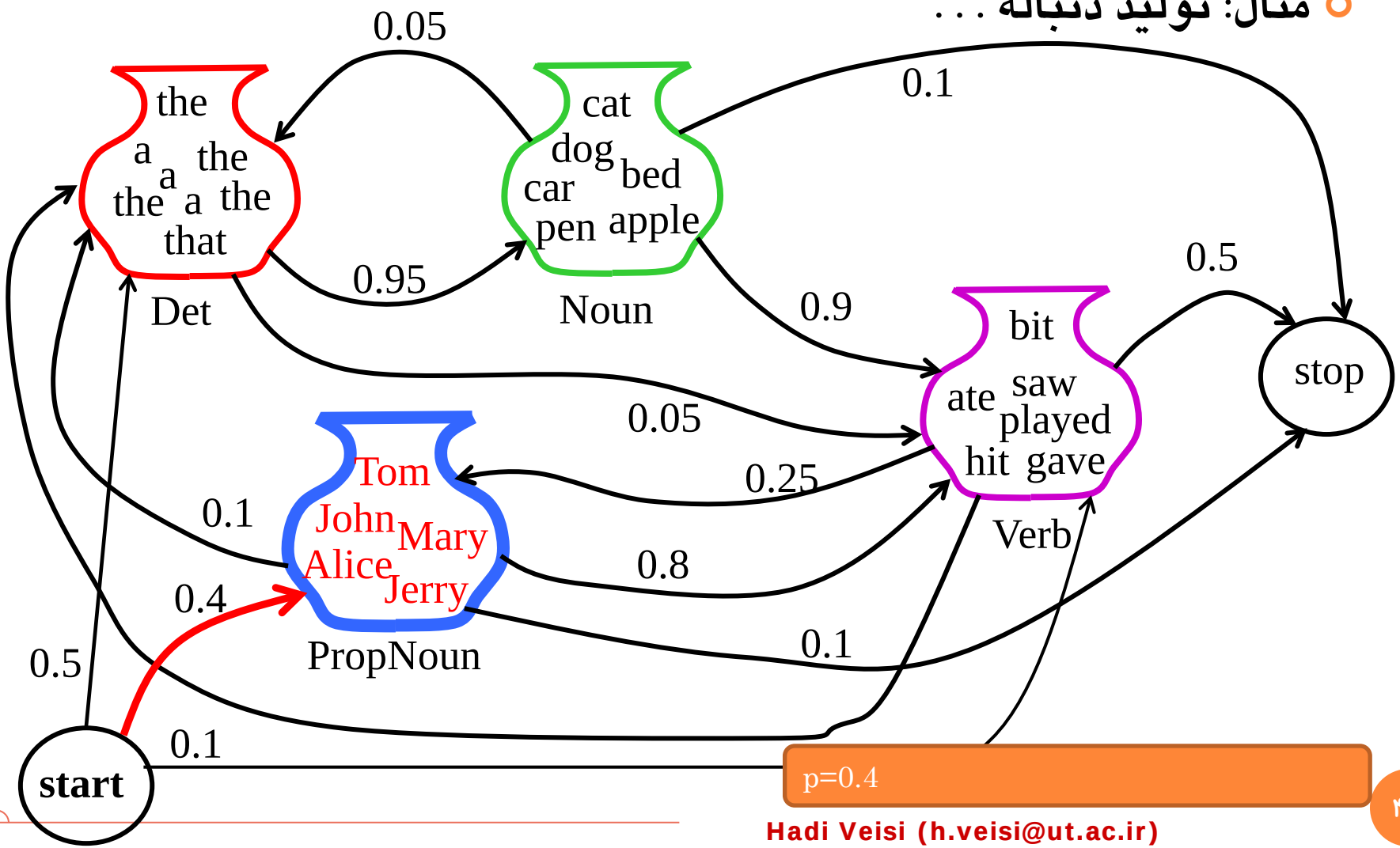
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال: تولید دنباله . . .



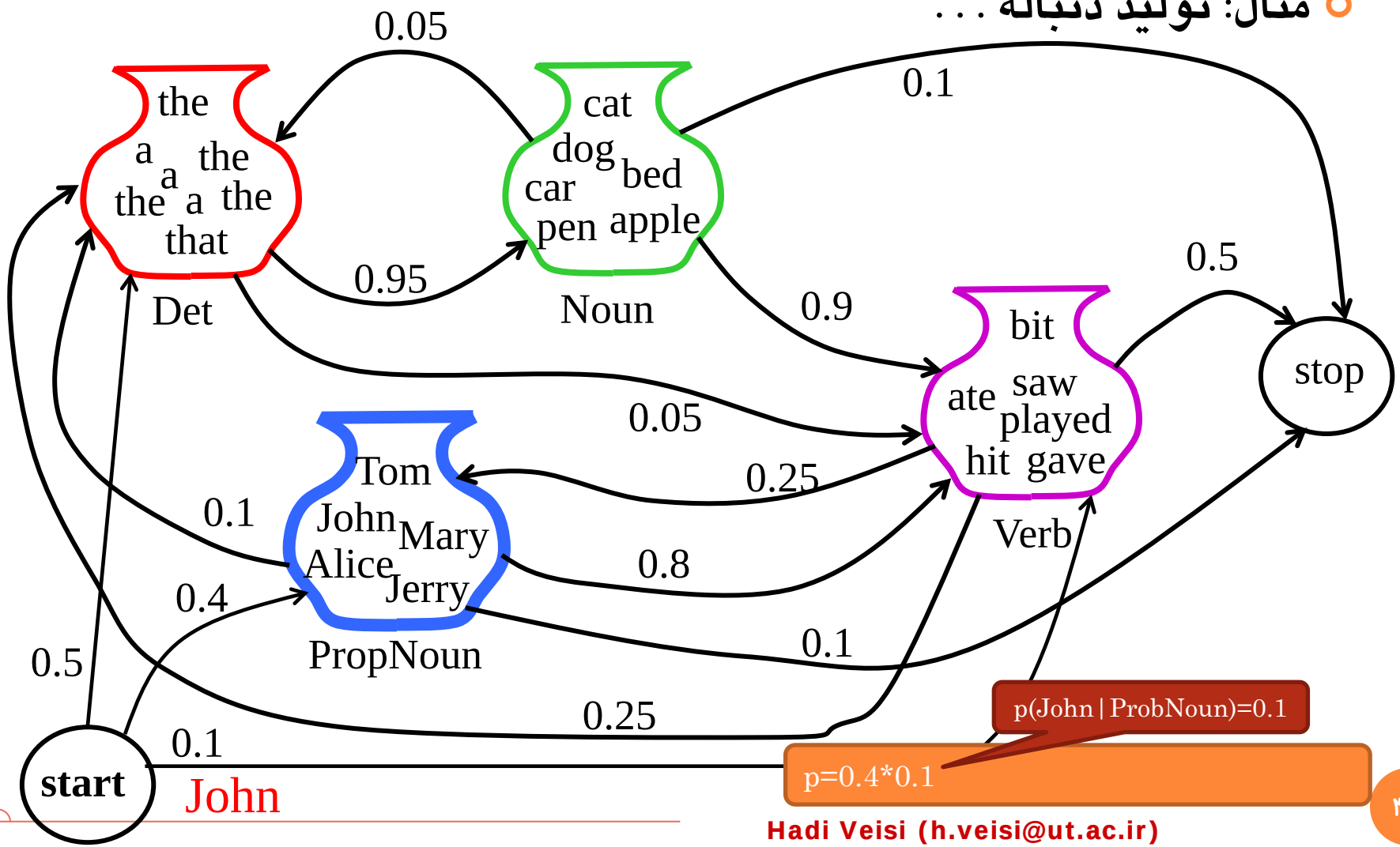
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال: تولید دنباله . . .



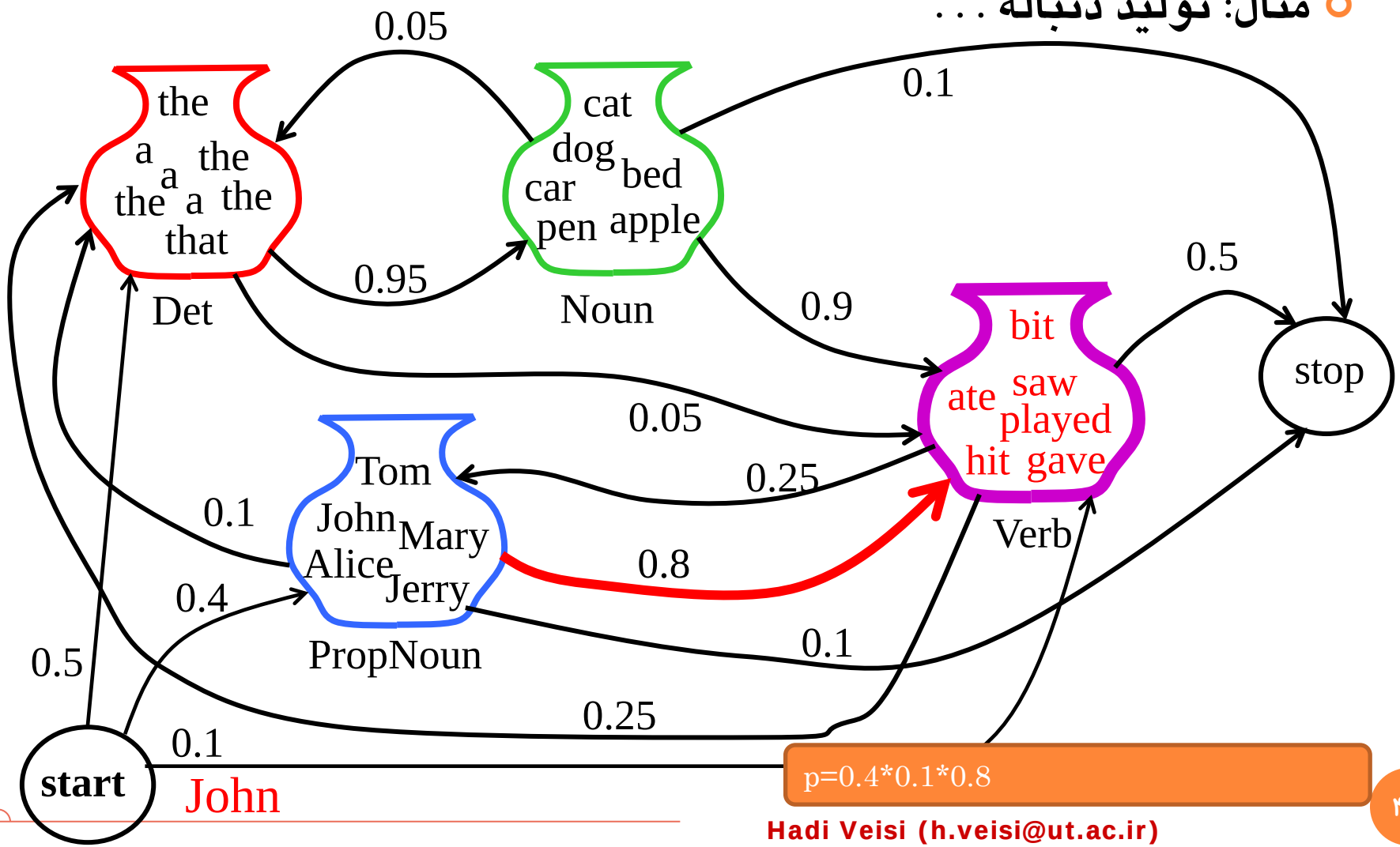
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال: تولید دنباله . . .



برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

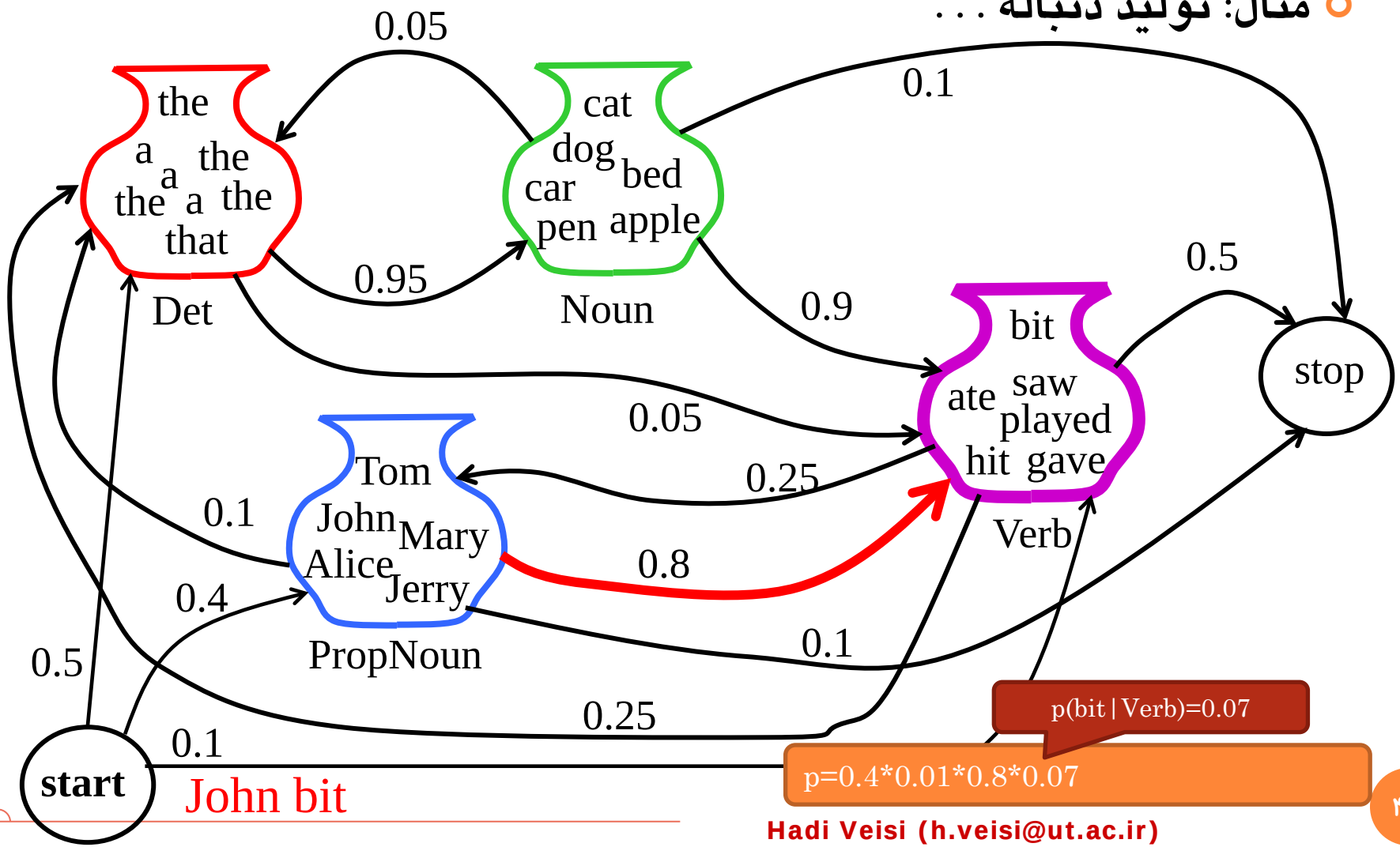
- مثال: تولید دنباله . . .



p=0.4*0.1*0.8

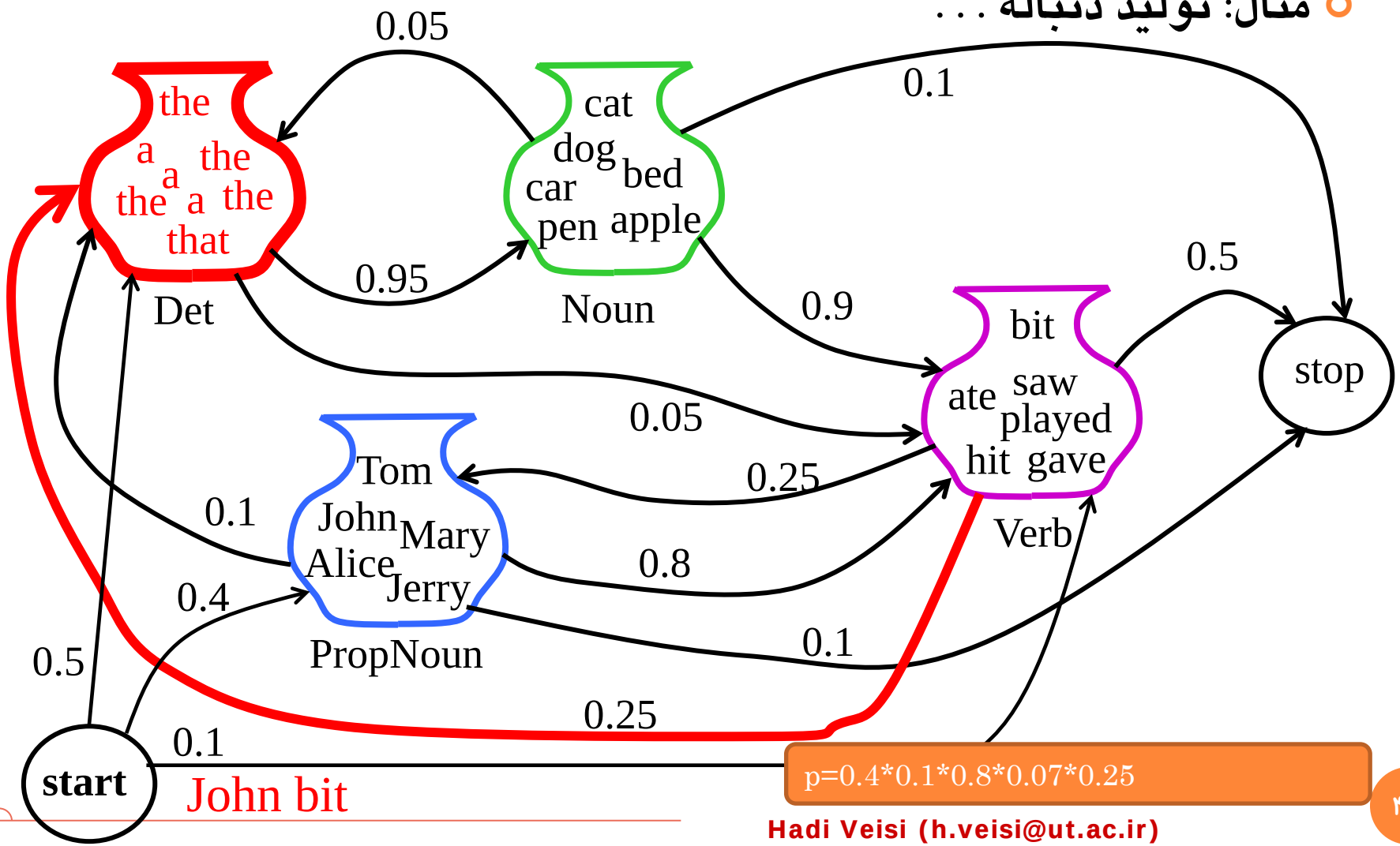
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال: تولید دنباله . . .



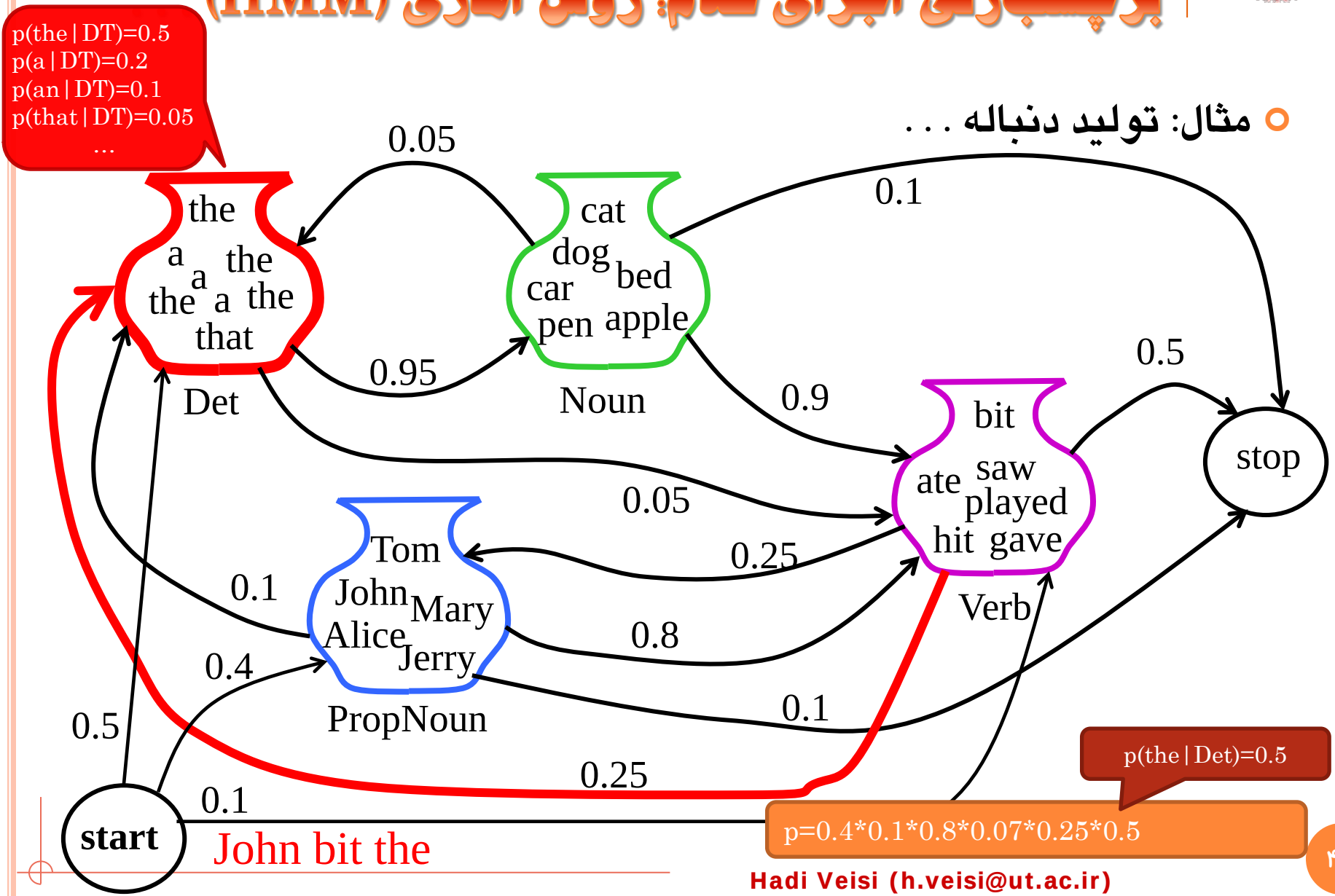
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال: تولید دنباله . . .



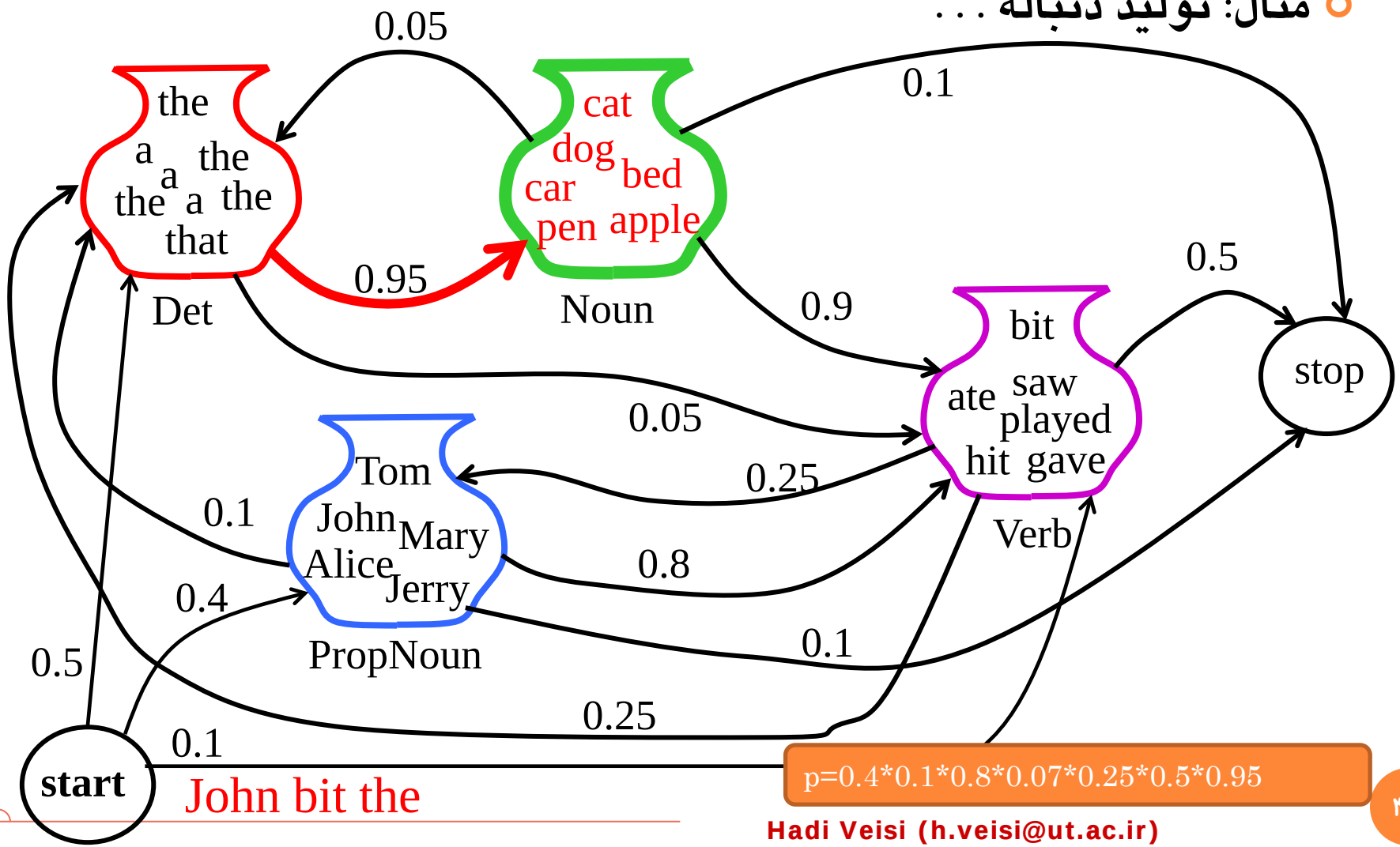
برچسب‌زنی اجزای کلام: روش آماری (HMM)

- مثال: تولید دنباله . . .



برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

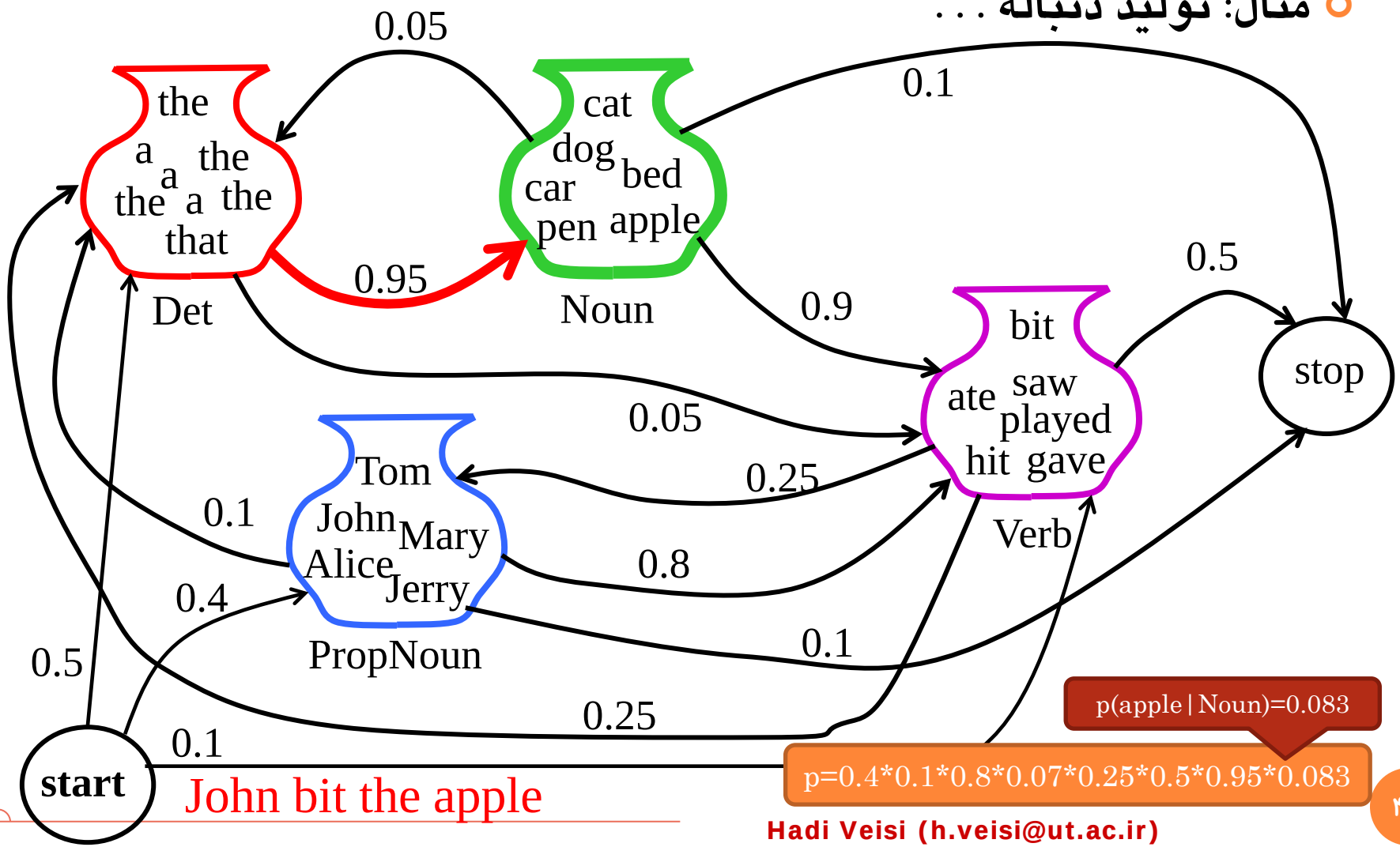
- مثال: تولید دنباله . . .



p=0.4*0.1*0.8*0.07*0.25*0.5*0.95

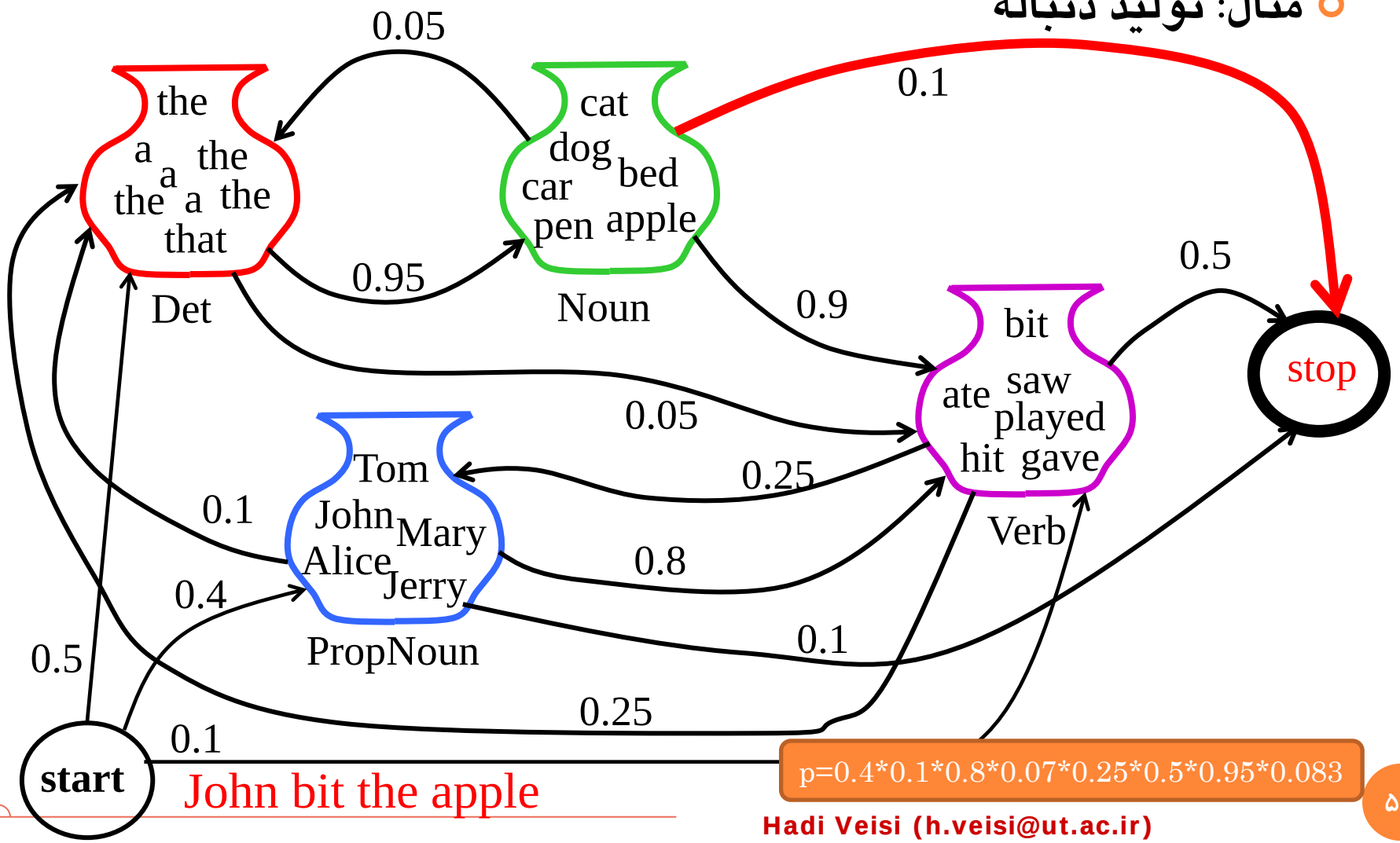
برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال: تولید دنباله . . .



برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- مثال: تولید دنباله



John bit the apple

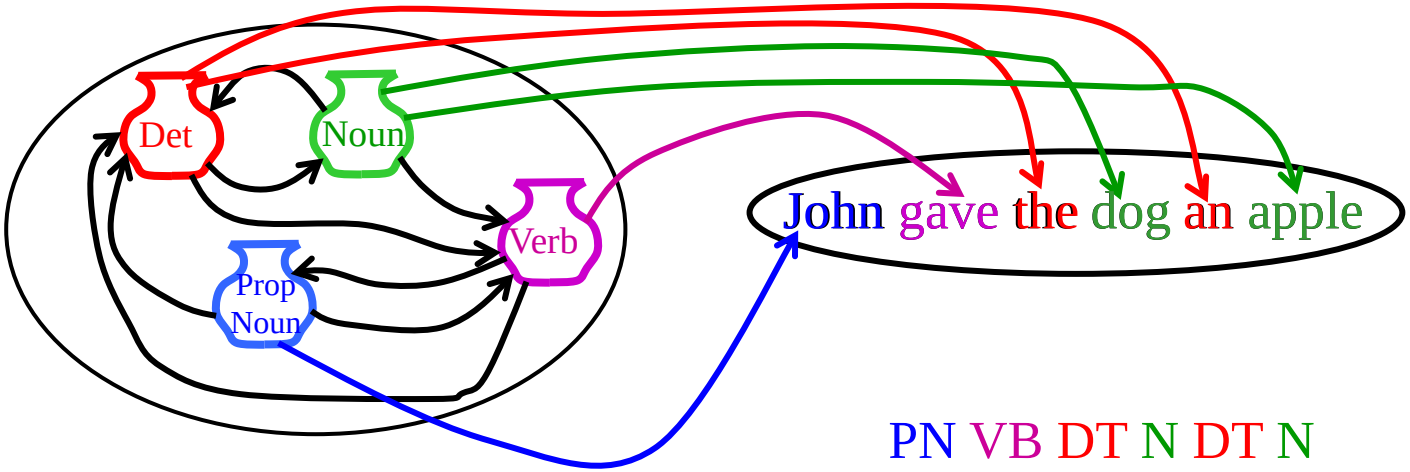
p=0.4*0.1*0.8*0.07*0.25*0.5*0.95*0.083

برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- **برچسب‌زنی اجزای کلام با روش HMM**
  - **فاز آموزش**
    - در نظر گرفتن یک واژگان با M کلمه و مجموعه برچسب‌های ممکن (N برچسب)
    - در نظر گرفتن یک HMM با تعداد حالت‌های برابر با تعداد برچسب‌ها (N حالت)
    - محاسبه احتمال‌های مدل با استفاده از یک پیکره متنی دارای برچسب اجزای کلام برای کلمات
      - احتمال اولیه حالت‌ها (N مقدار: هر حالت یک عدد)
      - احتمال انتقال از یک حالت (برچسب) به حالت دیگر (یک ماتریس N*N)
      - احتمال داشتن هر برچسب برای هر کلمه (یک ماتریس M*N)
  - **فاز آزمون (استفاده)**
    - دریافت یک دنباله از کلمات
    - یافتن بهترین برچسب‌های مرتبط= بهترین دنباله حالت در مدل HMM

برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- برچسب‌زنی اجزای کلام با روش HMM



برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

- یافتن دنباله حالت‌های (برچسب‌های) بهینه برای یک دنباله از کلمات
  - با داشتن دنباله مشاهده (کلمات) $O=O_1O_2\ldots O_T$ و مدل مخفی کارکوف λ، چگونه می‌توان بهترین دنباله حالت‌های (برچسب‌ها) $Q=q_1q_2\ldots q_T$ که متناسب با مشاهده است، را بدست آورد؟

- مساله دیکدینگ در HMM
- راه‌حل کامل: بررسی تمام دنباله حالت‌های ممکن و انتخاب بهترین آنها
  - بسیار زمان‌بر، از مرتبه $O(TN^T)$ که N تعداد حالت‌ها (برچسب‌ها) و T طول دنباله مشاهده‌ها (کلمات) است

- محاسبه محتمل‌ترین دنباله از برچسپ‌ها
  - ساده‌ترین روش: درنظر گرفتن تمام دنباله‌های محتمل و محاسبه احتمال هر یک به روش بیان شده (Brute Force Search)
  - با فرض داشتن N برچسب و T کلمه، حداکثر N^T دنباله از برچسب‌ها تولید می‌شود.
    - محاسبات بسیار زیاد
  - روش‌های رایج
    - مدل مخفی مارکوف (HMM: Hidden Markov Model)
    - میدان تصادفی شرطی (CRF: Conditional Random Field)

- راه حل بهینه: الگوریتم ویتربی (Viterbi)
  - یک روش برنامه نویسی پویا (Dynamic Programming)
    - مشابه الگوریتم Minimum Edit Distance

برچسب‌زنی اجزای کلام: روش آماری (HMM) . . .

الگوریتم ویتربی

function VITERBI(*observations* of len T,*state-graph* of len N) **returns** *best-path*

create a path probability matrix $viterbi[N+2,T]$

for each state s **from** 1 **to** N **do** ;initialization step

$viterbi[s,1] \leftarrow a_{0,s} * b_s(o_1)$

$backpointer[s,1] \leftarrow 0$

for each time step t **from** 2 **to** T **do** ;recursion step

for each state s **from** 1 **to** N **do**

$$viterbi[s,t] \leftarrow \max_{s'=1}^{N} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$$

$$backpointer[s,t] \leftarrow \operatorname{argmax}_{s'=1}^{N} viterbi[s',t-1] * a_{s',s}$$

$$viterbi[q_F,T] \leftarrow \max_{s=1}^{N} viterbi[s,T] * a_{s,q_F} \quad \text{; termination step}$$

$$backpointer[q_F,T] \leftarrow \operatorname{argmax}_{s=1}^{N} viterbi[s,T] * a_{s,q_F} \quad \text{; termination step}$$

return the backtrace path by following backpointers to states back in time from $backpointer[q_F,T]$

ماتریس احتمال‌ها

ماتریس دنباله حالت‌ها

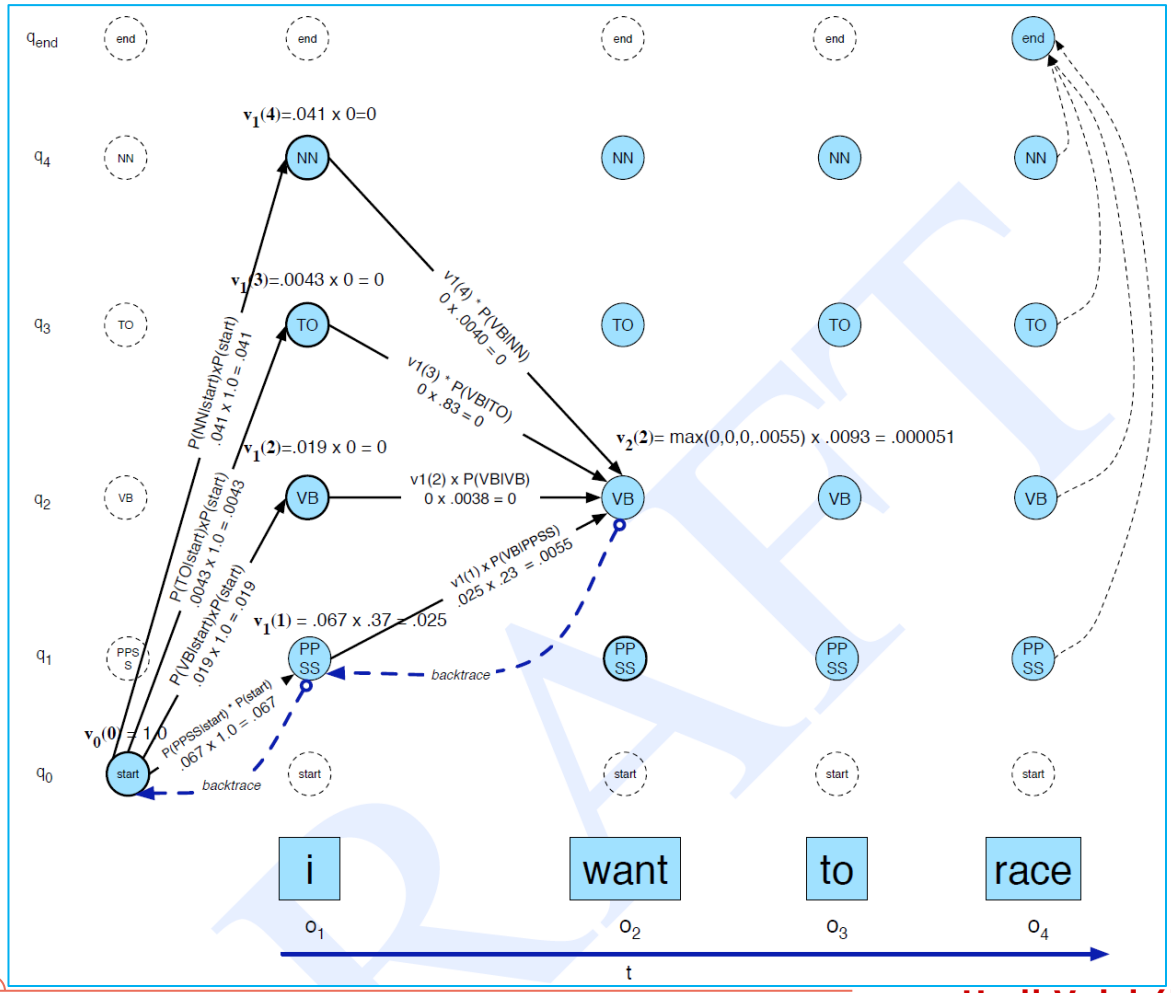
احتمال اولیه حالت s

احتمال اینکه کلمه o_t دارای حالت s باشد

برچسب‌زنی اجزای کلام: روش آماری (HMM)

الگوریتم ویتربی: مثال

- ۴ برچسب و ۴ کلمه



Transition probabilities: $P(t_i|t_{i-1})$

	VB	TO	NN	PPSS
start	0.019	0.0043	0.041	0.067
VB	0.0038	0.0345	0.047	0.070
TO	0.83	0	0.00047	0
NN	0.0040	0.016	0.087	0.0045
PPSS	0.23	0.00079	0.0012	0.00014

Observation likelihoods: $P(w_i|t_i)$

	I	want	to	race
VB	0	0.0093	0	0.00012
TO	0	0	0.99	0
NN	0	0.000054	0	0.00057
PPSS	0.37	0	0	0

برچسب‌زنی اجزای کلام: روش مبتنی بر تبدیل

- روش مبتنی بر تبدیل (Transformation-Based Tagging)
  - بر اساس ایده یادگیری مبتنی بر تبدیل (TBL: Transformation-Based Learning)
  - ترکیب روش‌های مبتنی بر قاعده و روش آماری
    - مبتنی بر قاعده: استفاده از قواعد برای شناسایی برچسب‌های نادرست
    - آماری: یادگیری قوانین به صورت خودکار از روی داده با روش‌های آماری
  - الگوریتم
    - انتساب محتمل‌ترین برچسب به کلمه (استفاده از واژگان)
    - تغییر برچسب بر اساس قوانین
      - *"if word-1 is a DT and word is a VRB then change the tag to NN"*
    - تکرار الگوریتم (تا رسیدن به شرط توقف)
  - ساخت قواعد بر اساس قالب‌های (templates) مشخص

The preceding (following) word is tagged **z**.
The word two before (after) is tagged **z**.
One of the two preceding (following) words is tagged **z**.
One of the three preceding (following) words is tagged **z**.
The preceding word is tagged **z** and the following word is tagged **w**.
The preceding (following) word is tagged **z** and the word two before (after) is tagged **w**.

#	Change tags From	To	Condition	Example
1	NN	VB	Previous tag is TO	to/TO race/NN → VB
2	VBP	VB	One of the previous 3 tags is MD	might/MD vanish/VBP → VB
3	NN	VB	One of the previous 2 tags is MD	might/MD not reply/NN → VB
4	VB	NN	One of the previous 2 tags is DT	
5	VBD	VBN	One of the previous 3 tags is VBZ	

برچسب‌زنی اجزای کلام: ارزیابی

- معیار
  - درصد کلماتی که درست برچسب زده شده است
  - محاسبه کارایی روی مجموعه آزمون (Test Set)
    - برابر با ۱۰ تا ۲۰ درصد پیکره دارای برچسب

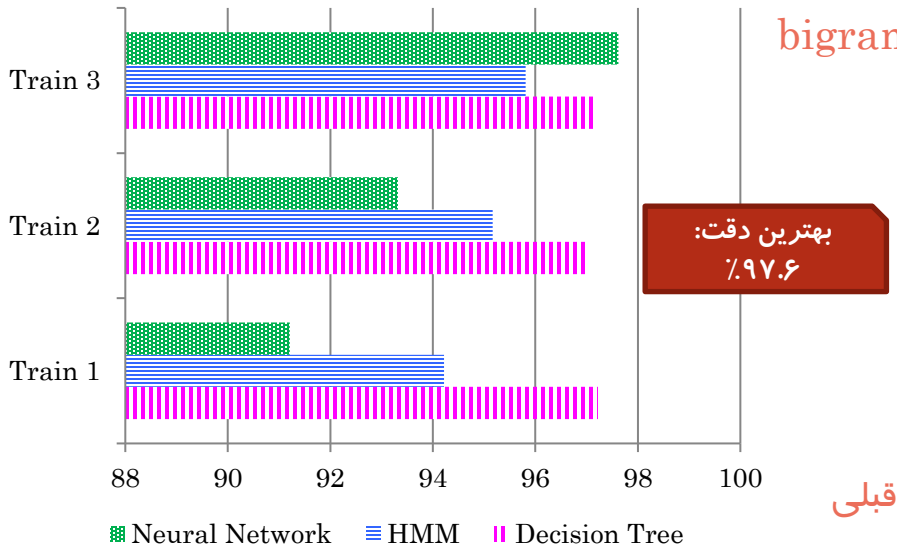
$$Acc = \frac{\text{تعداد کلمات با برچسب درست}}{\text{تعداد کل کلمات}} \times 100$$

- کارایی حدودی (برخی از روش‌ها)
  - برچسب زدن بر اساس محتمل‌ترین برچسب
    - دقت کلی:۹۰٪
    - دقت روی کلمات ناشناس: ۵۰٪
  - روش HMM (با احتمال‌های Trigram)
    - دقت کلی:۹۵٪
    - دقت روی کلمات ناشناس: ۵۵٪
  - انسان (کران بالا)
    - حدود ۹۸٪

برچسب‌زنی اجزای کلام: مثال (فارسی) . . .

- برچسب‌زنی اجزای کلام فارسی
  - پیکره: بیجن‌خان (تعداد برچسب‌ها = ۴۱)
  - روش‌ها
    - شبکه عصبی مصنوعی (Artificial Neural Network)
      - پرسپترون چند لایه (MLP)
    - درخت تصمیم (Decision Tree)
    - مدل مخفی مارکوف (HMM) – مبتنی بر bigram
  - اطلاعات (ویژگی‌ها) برای ANN و DT
    - کلمه فعلی
    - کلمه قبلی
    - برچسب کلمه قبلی
    - احتمال همه برچسب‌ها برای کلمه فعلی
    - احتمال همه برچسب‌ها برای کلمه برچسب قبلی

Name	# Samples
Train 1	200K
Train 2	1.0 M
Train 3	2.587 M
Test	10K



برچسب‌زنی اجزای کلام: مثال (فارسی)

- برچسب‌زنی اجزای کلام فارسی: قدرت تعمیم شبکه عصبی و HMM
  - پیکره ۱۰ میلیون کلمه ای بیجن خان با ۴۱ برچسب
    - آموزش = ۹۲٪، اعتبارسنجی = ٤٪ و آزمون = ٤٪ (حدود ۰.۸٪ کلمات OOV)
  - شبکه عصبی LSTM (یک طرفه و دوطرفه) و MLP
    - بردار ویژگی Word Vector
  - روش HMM دوتایی (Bigram)

MODEL	IV	OOV	Total
HMM	**96.30**	45.46	**95.82**
1-layer MLP	94.91	65.0	94.67
2-layer MLP	95.23	**66.0**	95.00
ULSTM	95.54	50.48	95.16
BLSTM	95.60	50.79	95.23