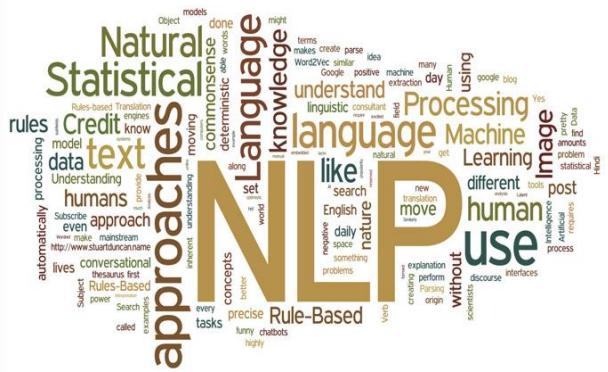


به نام خدا

گروه هوش مصنوعی، دانشکده مهندسی کامپیوتر

بخش چهارم

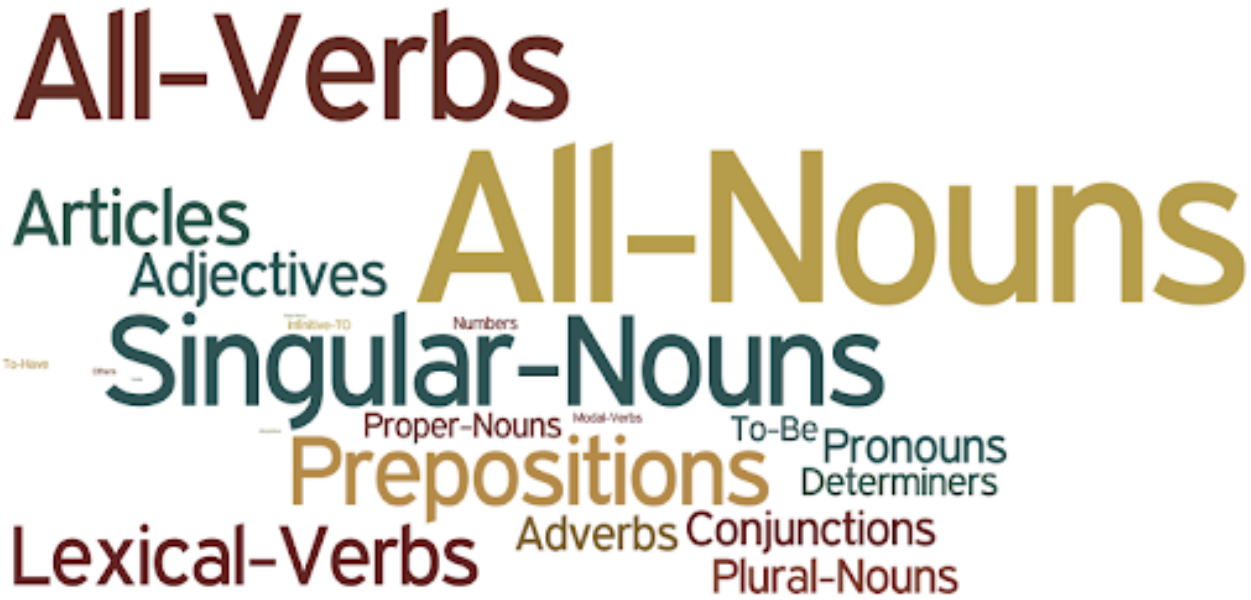
برچسب گذاری کلمات

Part-Of-Speech (POS)

حمیدرضا برادران کاشانی

سرفصل مطالب

- تاریخچه POS و کاربردهای آن
- تقسیم بندی POS: مجموعه باز و بسته
- انواع برچسب های گرامری
- ابهام در POS tagging
- ویژگی های رایج برای برچسب زنی
- روش های برچسب زنی و دقت آنها



Hamidreza Baradaran Kashani

تاریخچه برچسب گذاری کلمات

- قدمت موضوع به حدود ۳۰۰ سال قبل از میلاد برگردد (322-384 BCE)
  - ایده Parts-Of-Speech عنوان شد که به نوعی بیانگر:
    - lexical categories, word classes, tags, POS

- در ادامه Dionysius Thrax ۲۰۰ سال پس از شروع مساله، تعداد هشت برچسب را برای کلمات به عنوان POS معرفی کرد.
  - البته این برچسب ها با آنچه که ما تحت عنوان گرامر در مدرسه یادگرفته ایم قدری متفاوت است.

- Thrax: noun, verb, article, adverb, preposition, conjunction, participle, pronoun
- School grammar: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection

Hamidreza Baradaran Kashani

کاربرهای POS

- آنالیز احساسات
  - در بعضی از کاربردهای طبقه بندی متون ما تنها نیاز به کلماتی با برچسب خاص داریم.
  - مثلا برای آنالیز احساسات به کلماتی با برچسب صفت نیاز بیشتری است و سایر کلمات تاثیر کمتری دارند.
- تبدیل متن به گفتار (TTS)
  - بعضی کلمات دارای تلفظهای مختلفی با برچسبهای گوناگون هستند مثلا کلمه Live یا lead
- سیستمهای بازیابی اطلاعات (IR)
  - نقش POS در WSD یا Word Sense Disambiguation یا رفع ابهام از معنی کلمه

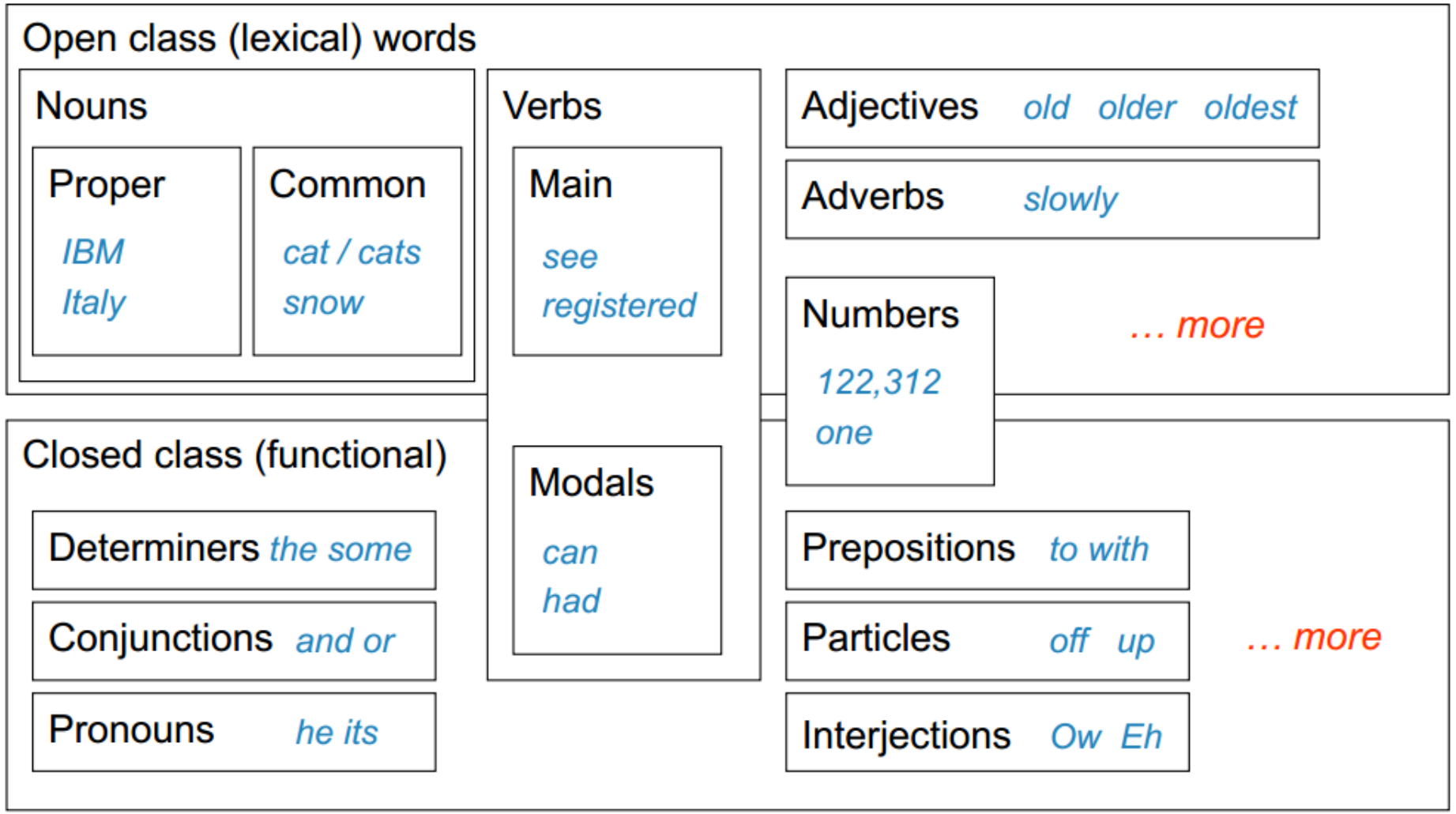
Hamidreza Baradaran Kashani

کاربرهای POS

اهمیت POS در بازیابی اطلاعات ؟

Hamidreza Baradaran Kashani

تقسیم بندی برچسبهای POS: مجموعه باز و بسته



ıradaran Kashani

تقسیم بندی برچسبهای POS: مجموعه باز و بسته

❖ مجموعه بسته

- determiners: ***a, an, the***
- pronouns: ***she, he, I***
- prepositions: ***on, under, over, near, by, …***

❖ مجموعه باز

- Nouns, Verbs, Adjectives, Adverbs.

Hamidreza Baradaran Kashani

انواع برچسب های گرامری

❖ Thrax POS Tags

❖ PenTree Bank POS Tags

❖ این tagger نسبت به Thrax از تعداد تگ های به مراتب بیشتری برخوردار است و در اکثر tagger ها از این برچسب ها استفاده می شود.

Hamidreza Baradaran Kashani

Thrax POS Tags

Tag	اصطلاح فارسی	مثال
verb	فعل	eat
noun	اسم	Jurafsky
article	حرف تعریف	a, an, the
adverb	قید	slowly
preposition	حرف اضافه	about, in
conjunction	حرف ربط	and, or, but
participle	صفت مفعولی	written, teaching
pronoun	ضمیر	i, he, it, they

Hamidreza Baradaran Kashani

PenTree Bank POS Tags

ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative

PenTree Bank POS Tags

9.	JJS	Adjective, superlative	18.	PRP	Personal pronoun
10.	LS	List item marker	19.	PRP$	Possessive pronoun
11.	MD	Modal	20.	RB	Adverb
12.	NN	Noun, singular or mass	21.	RBR	Adverb, comparative
13.	NNS	Noun, plural	22.	RBS	Adverb, superlative
14.	NNP	Proper noun, singular	23.	RP	Particle
15.	NNPS	Proper noun, plural	24.	SYM	Symbol
16.	PDT	Predeterminer	25.	TO	*to*
17.	POS	Possessive ending	26.	UH	Interjection

Hamidreza Baradaran Kashani

PenTree Bank POS Tags

27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Hamidreza Baradaran Kashani

ابهام در POS tagging

- بیشتر کلمات دارای بیش از یک برچسب هستند.
- مثال ۱:
  - The *back* door = JJ
  - On my *back* = NN
  - Promised to *back* the bill = VB
- مثال ۲:

Bill	saw	that	man	yesterday
NNP	NN	DT	NN	NN
VB	VB(D)	IN	VB	NN

هدف ما پیدا کردن برچسب صحیح کلمه با توجه به متن خاصی است که کلمه در آن قرار دارد.

Hamidreza Baradaran Kashani

ویژگی های رایج برای برچسب زنی

- برچسب های شایع کلمه (کلمه با احتمال بیشتر چه برچسبی دارد)
- اطلاعات کلمه مجاور
- پیشوند (Prefixes)

  - unfathomable: un → JJ
  - dissimilar: dis → JJ

- پسوند (Suffixes)

  - colorful: ful → JJ
  - Importantly: ly → RB

Hamidreza Baradaran Kashani

ویژگی های رایج برای برچسب زنی

w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i contains a number
w_i contains an upper-case letter
w_i contains a hyphen
w_i is all upper case
w_i's word shape
w_i's short word shape
w_i is upper case and has a digit and a dash (like *CFC-12*)
w_i is upper case and followed within 3 words by Co., Inc., etc.

Hamidreza Baradaran Kashani

روش های برچسب زنی و دقت آنها

❖ روش پایه

- ❖ استفاده از برچسب شایع هر کلمه بدون توجه به متن اطراف آن کلمه است.
- ❖ برچسب کلمات ناشناس را **Noun** قرار می دهند.
- ❖ روش راحت و نسبتا کارایی است (دقت حدود ۹۰٪) زیرا اکثر کلمات غیر مبهم هستند.

Hamidreza Baradaran Kashani

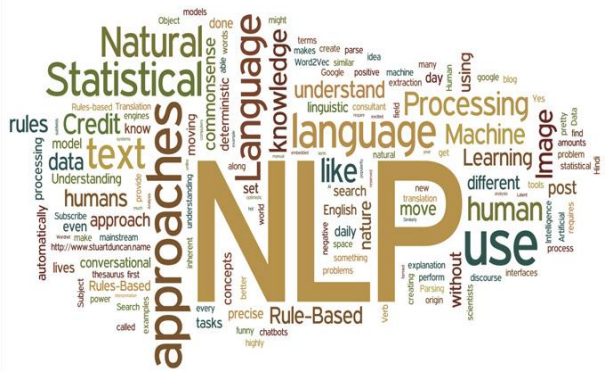
روش های برچسب زنی و دقت آنها

Rough accuracies:

- Most freq tag: ~90% / ~50%
- Trigram HMM: ~95% / ~55%
- Maxent P(t|w): 93.7% / 82.6%
- TnT (HMM++): 96.2% / 86.0%
- MEMM tagger: 96.9% / 86.9%
- Bidirectional dependencies: 97.2% / 90.0%
- Upper bound: ~98% (human agreement)

بیشتر خطاها به خاطر کلمات ناشناس (ناموجود در فاز آموزش) رخ می دهند

Hamidreza Baradaran Kashani



با تشکر از توجه شما