POS Tagging

PennTreebank

Hidden Markov model

Evaluation

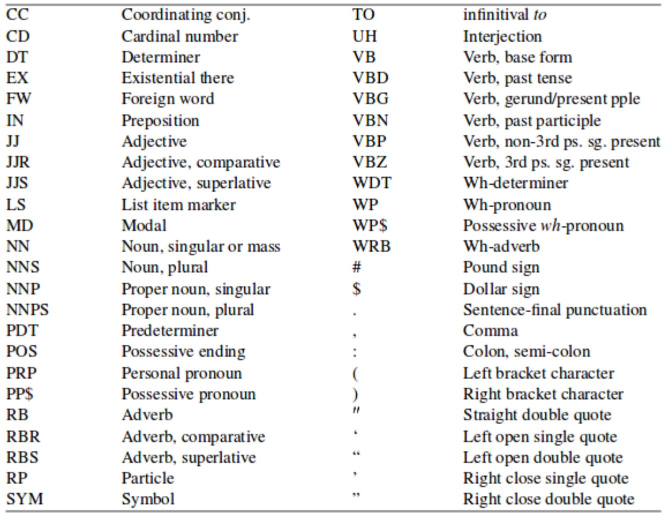
POS tagging

PennTreebank

- Created by University of Pennsylvania
- Eight-years project: 1989 – 1996
- 7 millions words of POS tagged texts
- POS tagset is based on Brown Corpus

POS tagging

PennTreebank



CC	Coordinating conj.	TO	infinitival *to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	’	Right close single quote
SYM	Symbol	”	Right close double quote

POS tagging

PennTreebank

- CC

  He bought a car and a house.

- CD

  Five years later, autocar will be popular.

- DT

  Pierre Vinken will join the board.

- EX

  There is no asbestos in our product now.

POS tagging

PennTreebank

- IN

  Mr Vinken is chairman of Elsevier N.V.

- JJ

  Rudolph Agnew was named an executive director.

- JJR

  The number of death was higher than expected

POS tagging

PennTreebank

- JJS

  The percentage of lung cancer appears to be highest.

- MD

  US should regulate the class of asbestos.

- NN

  It’s more than three times the expected number.

- NNS

  Portfolio managers expect further declines in interest rates.

POS tagging

PennTreebank

- NNP

  Alexis Sanchez joined Manchester United yesterday.

- NNPS

  … the Japan Automobile Dealers’ Association...

- POS

  … at Monday’s auction

POS tagging

PennTreebank

- PRP

  It expects to obtain regulatory approval.

- PP$

  Shareholders approve its acquisition by Royal Trustco Ltd.

- RB

  … depends heavily on creativity

- RBR

  … worked for the project for more than six years

POS tagging

PennTreebank

- RBS

  the most mundane aspect of its workers
- TO

  He decided to stay

POS tagging

PennTreebank

- VB

  … to return home
- VBD

  the executives joined Mayor William
- VBG

  … before boarding the buses again
- VBN

  A buffet breakfast was held in the museum

POS tagging

PennTreebank

- VBP

  Plans that give advertisers disscount
- VBZ

  The plan is not an attempt
- WDT

  a project that did not include Seymor
- WP

  who couldn’t be reach for comment

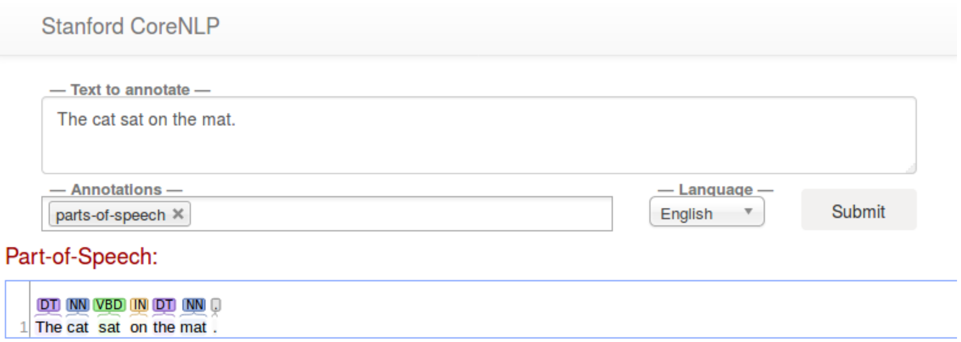
POS tagging

PennTreebank

- WRB

  where employees are assigned lunch partners

corenlp.run



Stanford CoreNLP

— Text to annotate —

The cat sat on the mat.

— Annotations —

parts-of-speech ✕

— Language —

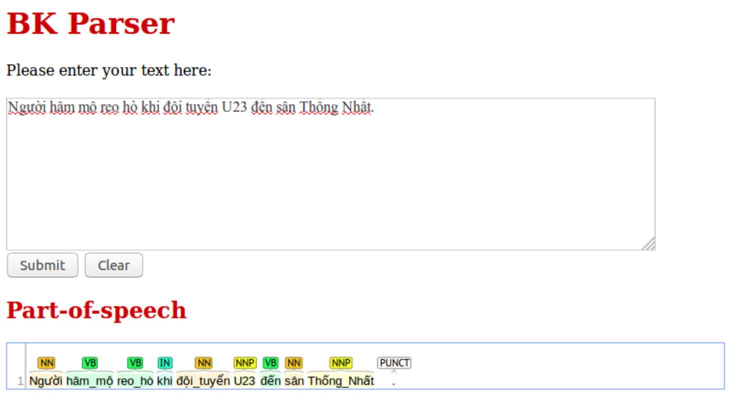
English

Submit

Part-of-Speech:

DT	NN	VBD	IN	DT	NN	.
The	cat	sat	on	the	mat	.

http://45.117.171.213/bknlptool/



BK Parser

Please enter your text here:

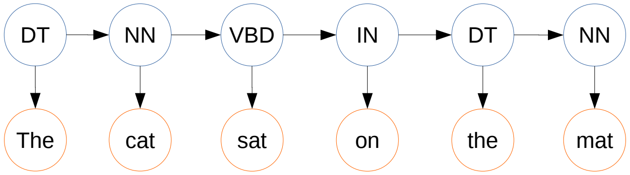
Người hâm mộ reo hò khi đội tuyển U23 đến sân Thống Nhất.

Part-of-speech

NN	VB	VB	IN	NN	NNP	VB	NN	NNP	PUNCT
Người	hâm_mộ	reo_hò	khi	đội_tuyển	U23	đến	sân	Thống_Nhất	.

POS tagging

Hidden Markov Models



POS tagging

Hidden Markov Models

- Transition probability

  $Pr(x_t = NN \mid x_{t-1} = DT)$

- Emission probabitlity

  $Pr(o_t = cat \mid x_t = NN)$

Hidden Markov Models

- Unsupervised parameter learning with MLE

  argmax$_{\textit{\textbf{theta}}}$ *Pr(**O, X** | **theta**)*

  Baum–Welch algorithm

- Decoding:

  argmax$_{\textit{X}}$ *Pr(**X** | **theta, O**)*

  Viterbi algorithm

Baum-Welch algorithm

- E step
  - Forward phase

$$\alpha_i(t) = P(o_1 o_2..\grave{o}_{t-1}, s_t = q_i|\lambda).$$

  - Backward phase

$$\beta_i(t) = P(o_{t+1} o_{t+2}..o_T, s_t = q_i|\lambda).$$

Baum-Welch algorithm

- M step

$$\gamma_i(t) = P(X_t = i|Y, \theta) = \frac{P(X_t = i, Y|\theta)}{P(Y|\theta)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)},$$

$$\xi_{ij}(t) = P(X_t = i, X_{t+1} = j|Y, \theta) = \frac{P(X_t = i, X_{t+1} = j, Y|\theta)}{P(Y|\theta)} = \frac{\alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})},$$

Viterbi decoding

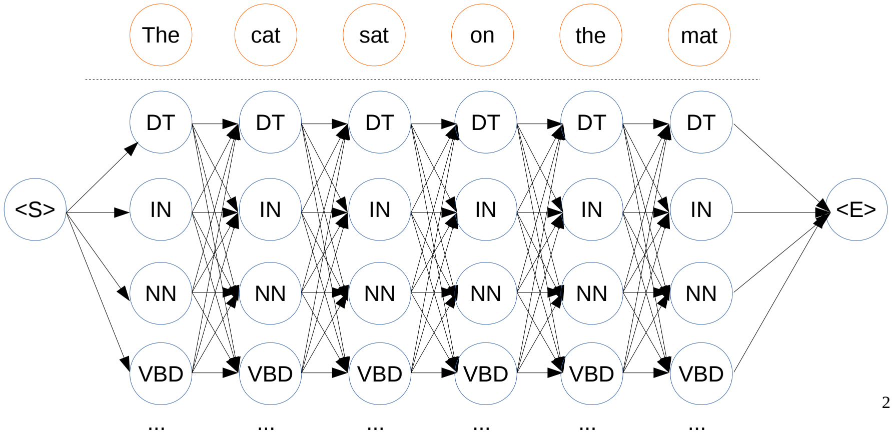
$$Best[i,t] = P(\hat{s_1}\hat{s_2}..\hat{s_{t-1}}, \hat{s}_t = q_i|o_1 o_2..o_t, \lambda).$$

$$Best[i,t] = max_j(Best[j,t-1] * a_{j,i} * b_{i,o_t})$$

$$Trace[i,t] = argmax_j(Best[j,t-1] * a_{j,i} * b_{i,o_t})$$

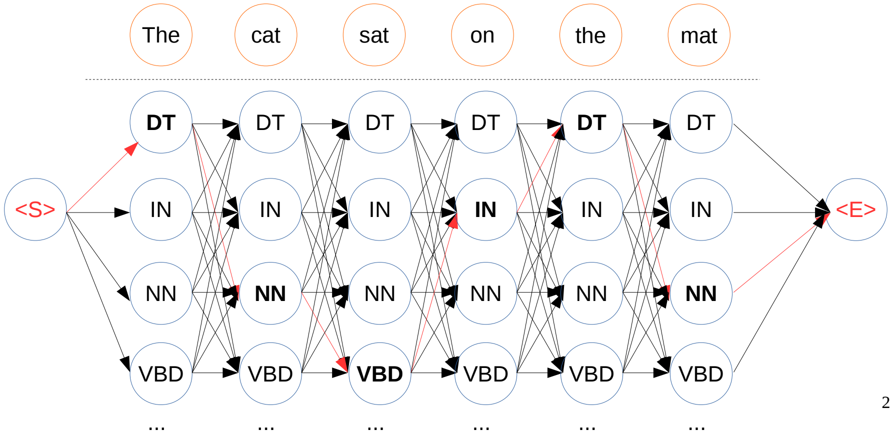
POS tagging
Viterbi decoding

$\mathit{argmax}_x\ P(\mathbf{X} \mid \mathbf{O}, \boldsymbol{theta})$



POS tagging
Viterbi decoding

$\mathit{argmax}_x\ P(\mathbf{X} \mid \mathbf{O}, \boldsymbol{theta})$



POS tagging
Supervised paramter estimation

- Transition probability

  $Pr(x_t=NN|x_{t-1}=DT)$

- Emission probabitlity

  $Pr(o_t=cat|x_t=NN)$

- Supervised parameter estimation

  $Pr(x_t=NN|x_{t-1}=DT)=(count(DT,NN)+1)/(count(DT)+L)$

  $Pr(o_t=cat|x_t=NN)=(count(cat,NN)+1)/(count(NN)+V)$

POS Tagging
Evaluation

- Comparing system output with golden annotations
- Datasets:

  Train: Used to train taggers

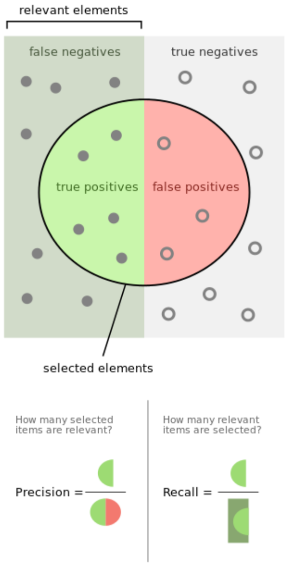
  Dev: Used to tune hyper-parameters

  Test: Used to test models

POS Tagging

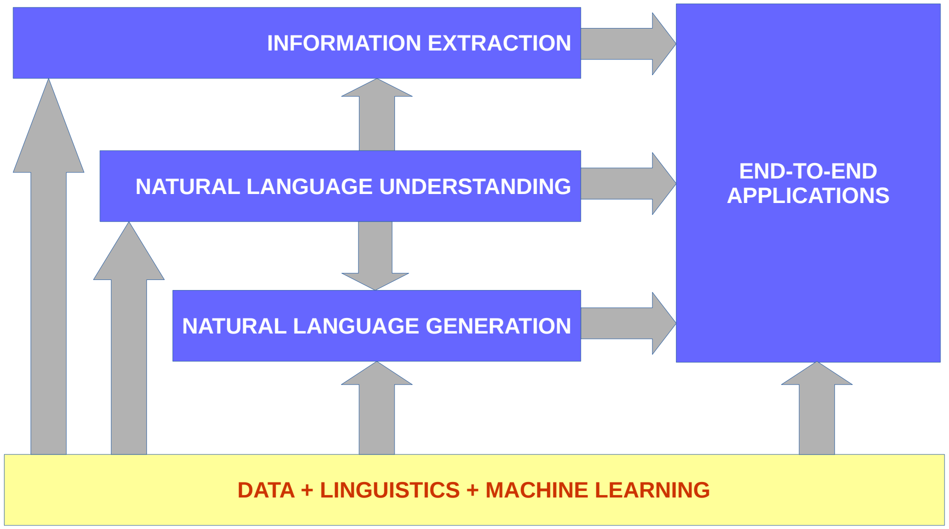
Evaluation

- Precision
- Recall
- $F_1 = 2PR / (P+R)$



Advanced topics in NLP

- Featured-based machine learning (e.g. SVMs, CRFs)
- Deep learning (e.g. word2vec, RNNs, CNNs, seq-2-seq)
- Transfer learning, reinforcement learning



INFORMATION EXTRACTION

NATURAL LANGUAGE UNDERSTANDING

NATURAL LANGUAGE GENERATION

END-TO-END APPLICATIONS

DATA + LINGUISTICS + MACHINE LEARNING

Q&A

hieunk@soict.hust.edu.vn