Natural Language Processing

Seminar 2

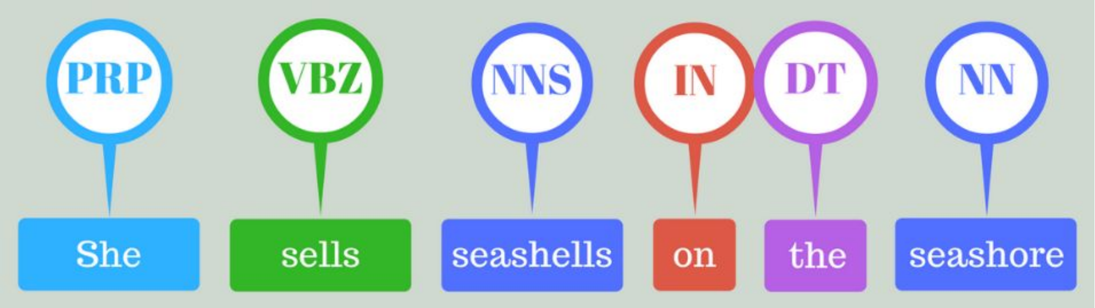
CMC MSU, February 18, 2017

Sequence models in NLP

Outline:

- Models Zoo
  - Hidden Markov Model
  - Maximum Entropy Markov Model
  - Linear-chain CRF
- Applied tasks
  - Features engineering for NER
  - POS-tagging in NLTK

Sequence modes in NLP

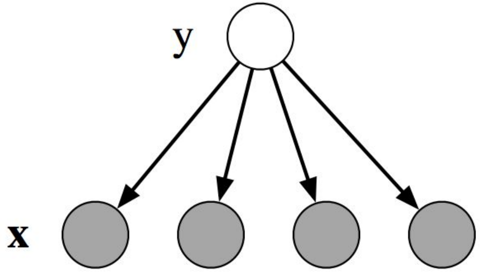


- Independent classifier for every position
- Graphical model
  - generative (HMM aka Naive Bayes)
  - discriminative (MEMM, CRF aka Logistic Regression)

Recap: Naive Bayes

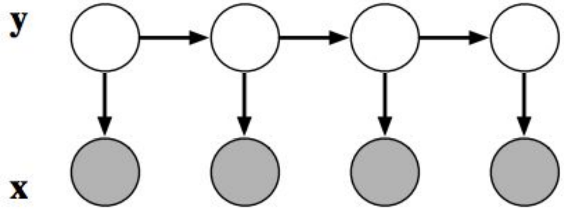
Model (**x** - feature vector, y - one label):

$$p(y, \mathbf{x}) = p(y) \prod_{k=1}^{K} p(x_k | y).$$



- Training: estimate probabilities by likelihood maximization
- Inference: y* = argmax p(y, x)

Hidden Markov Model

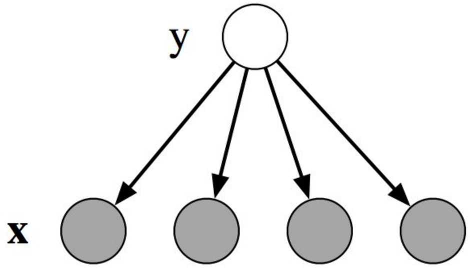


Model:

$$p(\mathbf{y},\mathbf{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t).$$

- Training: Baum-Welch algorithm
  - E-step: Forward-Backward (expectation over hidden variables)
  - M-stem: Likelihood maximization (update parameters)
- Inference (decoding): Viterbi algorithm

Recap: Logistic Regression (MaxEnt)



Model:

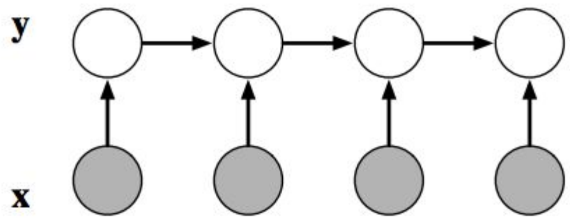
$$p(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\theta_y + \sum_{j=1}^{K} \theta_{y,j} x_j\right\}$$

In other notation:

$$p(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(y, \mathbf{x})\right\}$$

Training: conditional likelihood maximization (e.g. by SGD)

Maximum Entropy Markov Model



Model:

$$p_{\text{MEMM}}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1}, \mathbf{x})$$

$$p(y_t|y_{t-1}, \mathbf{x}) = \frac{1}{Z_t(y_{t-1}, \mathbf{x})} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

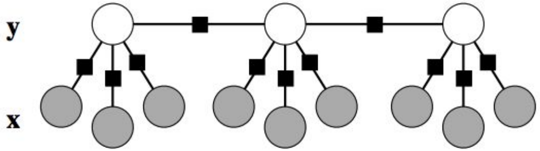
Training: convex optimization e.g. SGD + EM-algorithm

Inference (decoding): analogue to Viterbi algorithm

Feature engineering

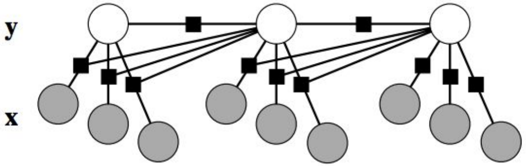
- Categorical features
- Label-observation features
- Edge-observation and node-observation features
- Features from different time stamps
- Boundary labels
- Features as backoff
- Unsupported features
- ...

Linear chain CRF



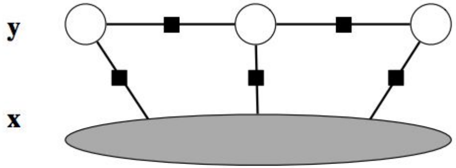
Model:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

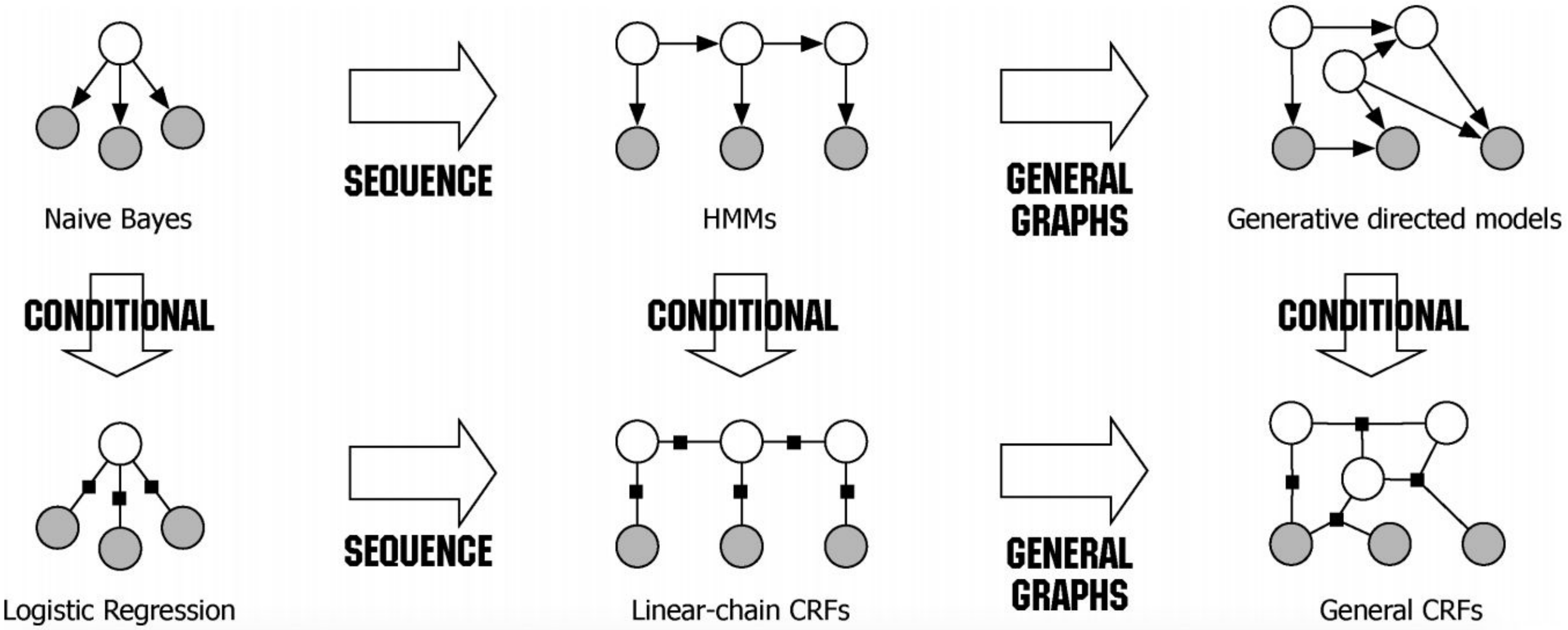


- Undirected graphical model
- Conditional probability from HMM is equal to CRF with particular choice of feature functions
- Inference: e.g. belief propagation
- General case:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{a=1}^{A} \Psi_a(\mathbf{y}_a, \mathbf{x}_a).$$



Models zoo summary



Common NLP sequence tasks

- Part-Of-Speech tagging (POS)
- Chunking (e.g. noun groups)
- Named Entity Recognition (NER)
- Word Sense Disambiguation (WSD)
- Syntax (shallow parsing)
- Semantic Slot Filling
- ...

POS tags (Penn Treebank)

CC	Coordinating conjunction
CD	Cardinal number
CDT	Determiner
CEX	Existential there
CFW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WWP	Wh-pronoun
WRB	Wh-adverb

NER tags (CoNLL 2003 shared task)

$$\mathcal{Y} = \{\text{B-P}\textsc{er}, \text{I-PER}, \text{B-L}\textsc{oc}, \text{I-L}\textsc{oc}, \text{B-O}\textsc{rg}, \text{I-O}\textsc{rg}, \text{B-M}\textsc{isc}, \text{I-M}\textsc{isc}, \text{O}\}$$

U.N. official Ekeus heads for Baghdad.

PER, ORG, LOC, MISC labels + BIO-notation

Feature engineering

Table 2.2. A subset of observation functions $q_s(\mathbf{x}, t)$ for the CoNLL 2003 Engilsh named-entity data, used by Mccallum and Li [86].

W=v	$w_t = v$	$\forall v \in \mathcal{V}$
T=j	part-of-speech tag for w_t is j (as determined by an automatic tagger)	$\forall$POS tags j
P=I$-j$	w_t is part of a phrase with syntactic type j (as determined by an automatic chunker)	
Capitalized	w_t matches `[A-Z][a-z]+`	
Allcaps	w_t matches `[A-Z][A-Z]+`	
EndsInDot	w_t matches `[^\.]+.*\.`	
	w_t contains a dash	
	w_t matches `[A-Z]+[a-z]+[A-Z]+[a-z]`	
Acro	w_t matches `[A-Z][A-Z\\.]*\\.[A-Z\\.]*`	
Stopword	w_t appears in a hand-built list of stop words	
CountryCapital	w_t appears in list of capitals of countries	
$\vdots$	many other lexicons and regular expressions	
$q_k(\mathbf{x}, t+\delta)$ for all k and $\delta \in [-1, 1]$		

Implementation details

CRF++	http://crfpp.sourceforge.net/
MALLET	http://mallet.cs.umass.edu/
GRMM	http://mallet.cs.umass.edu/grmm/
CRFSuite	http://www.chokkan.org/software/crfsuite/
FACTORIE	http://www.factorie.cc

Table 5.1. Scale of typical CRF applications in natural language processing.

Task	Parameters	Observation Functions	# Sequences	# Positions	Labels	Time (s)
NP chunking	248471	116731	8936	211727	3	958s
NER	187540	119265	946	204567	9	4866s
POS tagging	509951	127764	38219	912344	45	325500s

Practice (next time)

- POS-taggers in NLTK
- Viterbi algorithm
- Language modeling