Information Extraction

Lecture 12 – More Machine Learning

CIS, LMU München
Winter Semester 2015-2016

Dr. Alexander Fraser, CIS

Administravia

- Today is the last lecture
- Please review all of the slides from the Vorlesung before next time
- Next time: Klausur review
- Time after that: Klausur (bring paper!)
- PLEASE MAKE SURE YOU ARE REGISTERED FOR THE KLAUSUR IN LSF!
  - Check again now!
- Also, if you are in the seminar, don't forget to register for that too (two registrations total!)

Lecture today

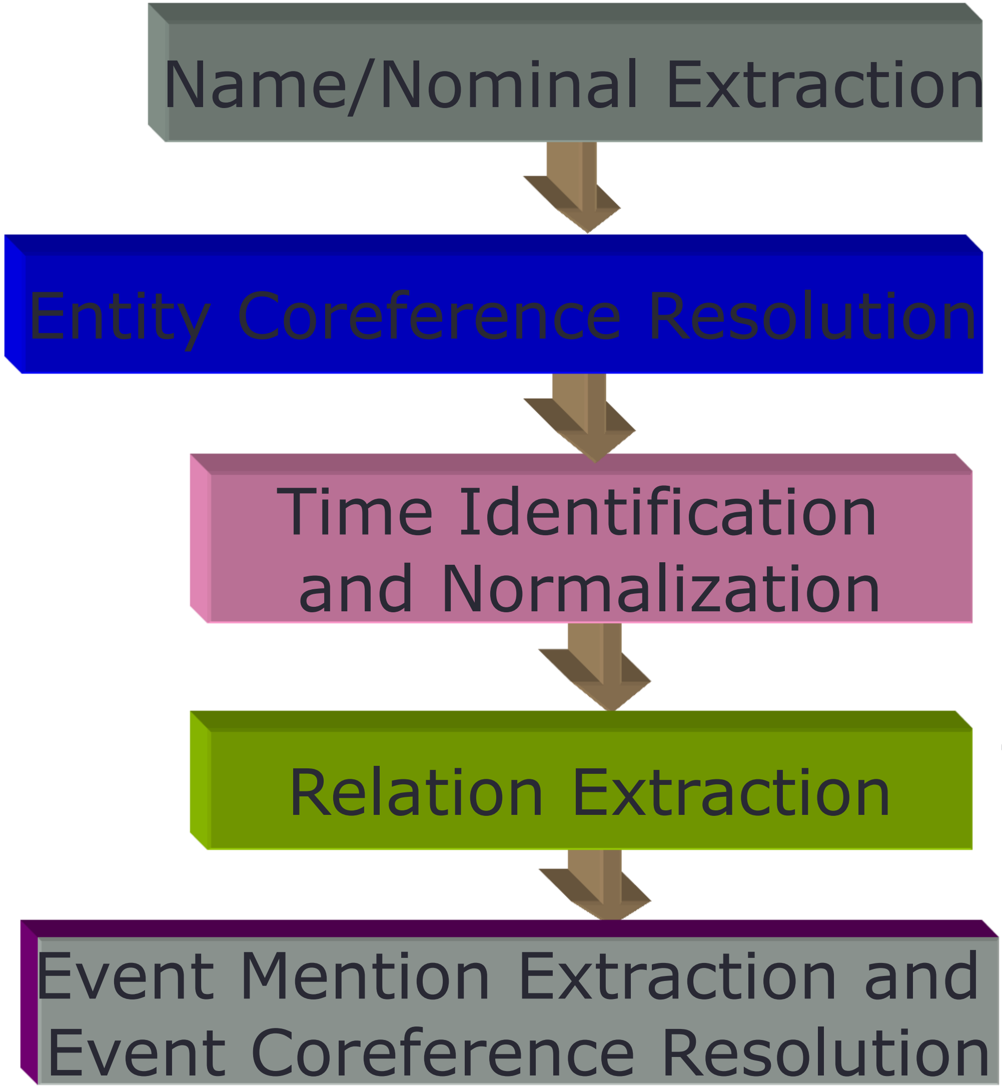
- Today we will go into more details in machine learning, particularly for NER
  - Also briefly discuss tagging different human languages
- We'll discuss the models which are used in Wapiti
  - Up until now we really only talked about the intuitions behind was is going on, rather than the real models (which are Maximum Entropy models, as we will see)
- In the last exercise, we'll look at sequence learning (rather than binary classification)
  - We'll also look briefly at regularization
- **Based on voting, the last exercise will be on Feb 3rd and 4th**

Supervised Learning based IE

- ‘Pipeline’ style IE
  - Split the task into several components
  - Prepare data annotation for each component
  - Apply supervised machine learning methods to address each component separately
  - Most state-of-the-art ACE IE systems were developed in this way
  - Provide great opportunity to applying a wide range of learning models and incorporating diverse levels of linguistic features to improve each component
  - Large progress has been achieved on some of these components such as name tagging and relation extraction

Slide from Heng Ji

Major IE Components



Name/Nominal Extraction

“Barry Diller”, “chief”

Entity Coreference Resolution

“Barry Diller” = “chief”

Time Identification and Normalization

Wednesday (2003-03-04)

Relation Extraction

“Vivendi Universal Entertainment” is *located in* “France”

Event Mention Extraction and Event Coreference Resolution

“Barry Diller” is *the person of the end-position event trigged by* “quit”

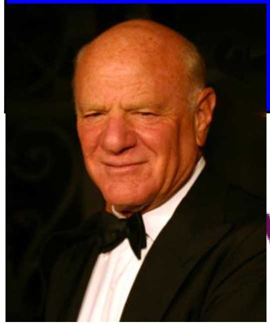
Slide from Heng Ji

IE Output

- (In this talk) Information Extraction (IE) =Identifying the instances of facts *names/entities , relations and events* from *semi-structured or unstructured* text; and convert them into structured representations (e.g. databases)

Barry Diller on Wednesday ***quit*** as chief of Vi **Vivendi Universal Entertainment**

Trigger	Quit (a "Personnel/End-Position" event)	
Arguments	Role = Person	Barry Diller
	Role = Organization	Vivendi Universal Entertainment
	Role = Position	Chief
	Role = Time-within	Wednesday (2003-03-04)



Slide modified from Heng Ji

Name Tagging

- Handcrafted systems
  - LTG
    - F-measure of 93.39 in MUC-7 (the best)
    - Ltquery, XML internal representation
    - Tokenizer, POS-tagger, SGML transducer
  - Nominator (1997)
    - IBM
    - Heavy heuristics
    - Cross-document co-reference resolution
    - Used later in IBM Intelligent Miner

Slide from Heng Ji

Name Tagging

- Handcrafted systems
  - LaSIE (Large Scale Information Extraction)
    - MUC-6 (LaSIE II in MUC-7)
    - Univ. of Sheffield's GATE architecture (General Architecture for Text Engineering)
    - JAPE language
  - FACILE (1998)
    - NEA language (Named Entity Analysis)
    - Context-sensitive rules
  - NetOwl (MUC-7)
    - Commercial product
    - C++ engine, extraction rules

Slide from Heng Ji

Automatic approaches

- Learning of statistical models or symbolic rules
  - Use of annotated text corpus
    - Manually annotated
    - Automatically annotated
- “BIO” tagging
  - Tags: Begin, Inside, Outside an NE
  - Probabilities:
    - Simple:
      - P(tag i | token i)
    - With external evidence:
      - P(tag i | token i-1, token i, token i+1)
- “OpenClose” tagging
  - Two classifiers: one for the beginning, one for the end

Slide from Heng Ji

Automatic approaches

- Decision trees
  - Tree-oriented sequence of tests in every word
    - Determine probabilities of having a BIO tag
  - Use training corpus
  - Viterbi, ID3, C4.5 algorithms
    - Select most probable tag sequence
  - SEKINE et al (1998)
  - BALUJA et al (1999)
    - F-measure: 90%

Slide from Heng Ji

Automatic approaches

- HMM
  - Markov models, Viterbi
  - Separate statistical model for each NE category + model for words outside NEs
  - Nymble (1997) / IdentiFinder (1999)
- Maximum Entropy (ME)
  - Separate, independent probabilities for every evidence (external and internal features) are merged multiplicatively
  - MENE (NYU - 1998)
    - Capitalization, many lexical features, type of text
    - F-Measure: 89%

Slide from Heng Ji

Automatic approaches

- Hybrid systems
  - Combination of techniques
    - IBM’s Intelligent Miner: Nominator + DB/2 data mining
  - WordNet hierarchies
    - MAGNINI et al. (2002)
  - Stacks of classifiers
    - Adaboost algorithm
  - Bootstrapping approaches
    - Small set of seeds
  - Memory-based ML, etc.

Slide from Heng Ji

NER in various languages

- Arabic
  - TAGARAB (1998)
  - Pattern-matching engine + morphological analysis
  - Lots of morphological info (no differences in orthographic case)
- Bulgarian
  - OSENOVA & KOLKOVSKA (2002)
  - Handcrafted cascaded regular NE grammar
  - Pre-compiled lexicon and gazetteers
- Catalan
  - CARRERAS et al. (2003b) and MÁRQUEZ et al. (2003)
  - Extract Catalan NEs with Spanish resources (F-measure 93%)
  - Bootstrap using Catalan texts

Slide modified from Heng Ji

NER in various languages

- Chinese & Japanese
  - Many works
  - Special characteristics
    - Character or word-based
    - No capitalization
  - CHINERS (2003)
    - Sports domain
    - Machine learning
    - Shallow parsing technique
  - ASAHARA & MATSMUTO (2003)
    - Character-based method
    - Support Vector Machine
    - 87.2% F-measure in the IREX (outperformed most word-based systems)

Slide from Heng Ji

NER in various languages

- Dutch
  - DE MEULDER et al. (2002)
    - Hybrid system
      - Gazetteers, grammars of names
      - Machine Learning Ripper algorithm
- French
  - BÉCHET et al. (2000)
    - Decision trees
    - Le Monde news corpus
- German
  - Non-proper nouns also capitalized
  - THIELEN (1995)
    - Incremental statistical approach
    - 65% of corrected disambiguated proper names

Slide from Heng Ji

NER in various languages

- Greek
    - KARKALETSIS et al. (1998)
        - English – Greek GIE (Greek Information Extraction) project
        - GATE platform
- Italian
    - CUCCHIARELLI et al. (1998)
        - Merge rule-based and statistical approaches
        - Gazetteers
        - Context-dependent heuristics
        - ECRAN (Extraction of Content: Research at Near Market)
        - GATE architecture
        - Lack of linguistic resources: 20% of NEs undetected
- Korean
    - CHUNG et al. (2003)
        - Rule-based model, Hidden Markov Model, boosting approach over unannotated data

Slide from Heng Ji

NER in various languages

- Portuguese
  - SOLORIO & LÓPEZ (2004, 2005)
    - Adapted CARRERAS et al. (2002b) Spanish NER
    - Brazilian newspapers
- Serbo-Croatian
  - NENADIC & SPASIC (2000)
    - Hand-written grammar rules
    - Highly inflective language
      - Lots of lexical and lemmatization pre-processing
    - Dual alphabet (Cyrillic and Latin)
      - Pre-processing stores the text in an independent format

Slide from Heng Ji

NER in various languages

- Spanish
  - CARRERAS et al. (2002b)
    - Machine Learning, AdaBoost algorithm
    - BIO and OpenClose approaches
- Swedish
  - SweNam system (DALIANIS & ASTROM, 2001)
    - Perl
    - Machine Learning techniques and matching rules
- Turkish
  - TUR et al (2000)
    - Hidden Markov Model and Viterbi search
    - Lexical, morphological and context clues

Slide from Heng Ji

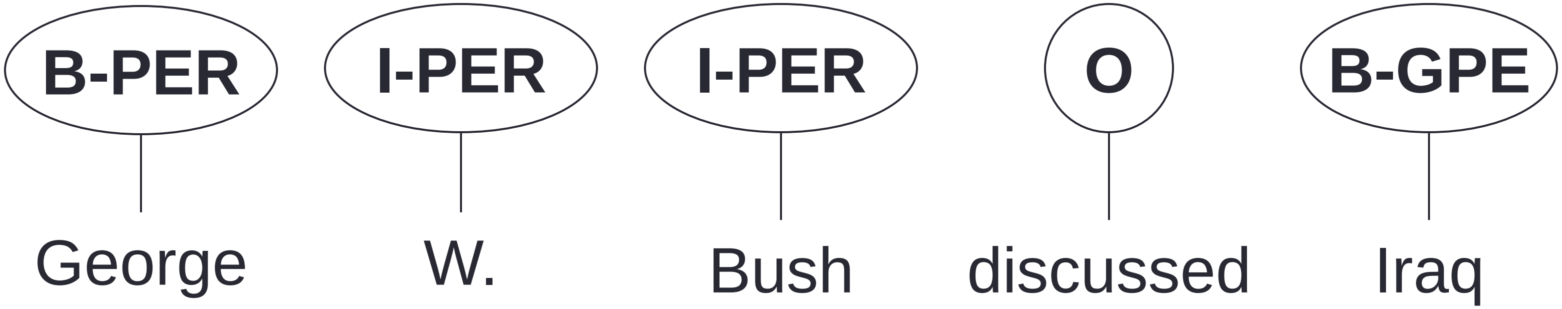
Name Tagging: Task

- Person (PER): named person or family
- Organization (ORG): named corporate, governmental, or other organizational entity
- Geo-political entity (GPE): name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)

<PER>George W. Bush</PER> discussed <GPE>Iraq</GPE>

- But also: Location, Artifact, Facility, Vehicle, Weapon, Product, etc.
- Extended name hierarchy, 150 types, domain-dependent (Sekine and Nobata, 2004)

- Convert it into a sequence labeling problem – “BIO” tagging:



B-PER I-PER I-PER O B-GPE

George W. Bush discussed Iraq

Slide from Heng Ji

Supervised Learning for Name Tagging

- Maximum Entropy Models (Borthwick, 1999; Chieu and Ng 2002; Florian et al., 2007)
- Decision Trees (Sekine et al., 1998)
- Class-based Language Model (Sun et al., 2002, Ratinov and Roth, 2009)
- Agent-based Approach (Ye et al., 2002)
- Support Vector Machines (Takeuchi and Collier, 2002)
- Sequence Labeling Models
  - Hidden Markov Models (HMMs) (Bikel et al., 1997; Ji and Grishman, 2005)
  - Maximum Entropy Markov Models (MEMMs) (McCallum and Freitag, 2000)
  - Conditional Random Fields (CRFs) (McCallum and Li, 2003)

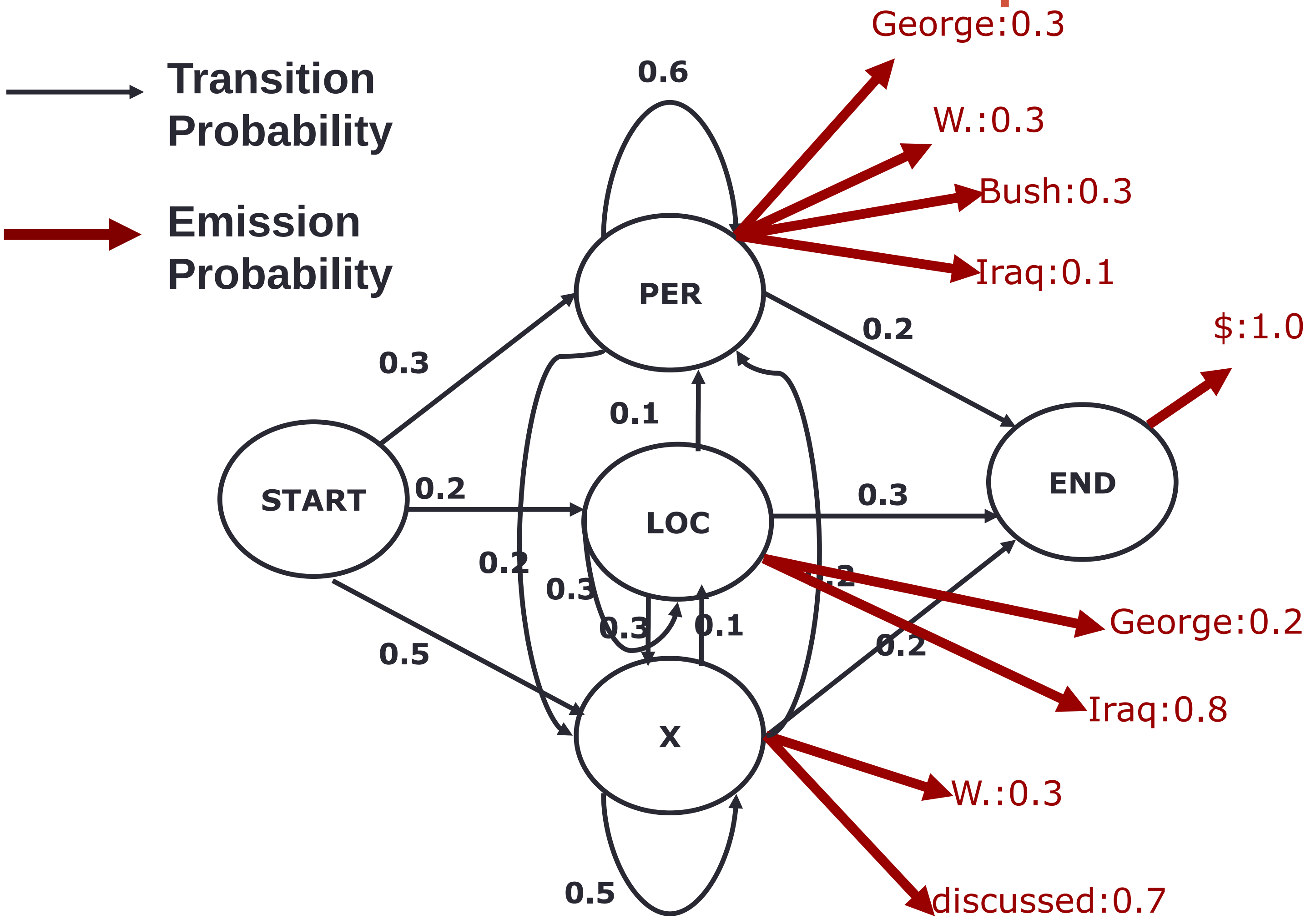
Slide from Heng Ji

Typical Name Tagging Features

- N-gram: Unigram, bigram and trigram token sequences in the context window of the current token
- Part-of-Speech: POS tags of the context words
- Gazetteers: person names, organizations, countries and cities, titles, idioms, etc.
- Word clusters: to reduce sparsity, using word clusters such as Brown clusters (Brown et al., 1992)
- Case and Shape: Capitalization and morphology analysis based features
- Chunking: NP and VP Chunking tags
- Global feature: Sentence level and document level features. For example, whether the token is in the first sentence of a document
- Conjunction: Conjunctions of various features

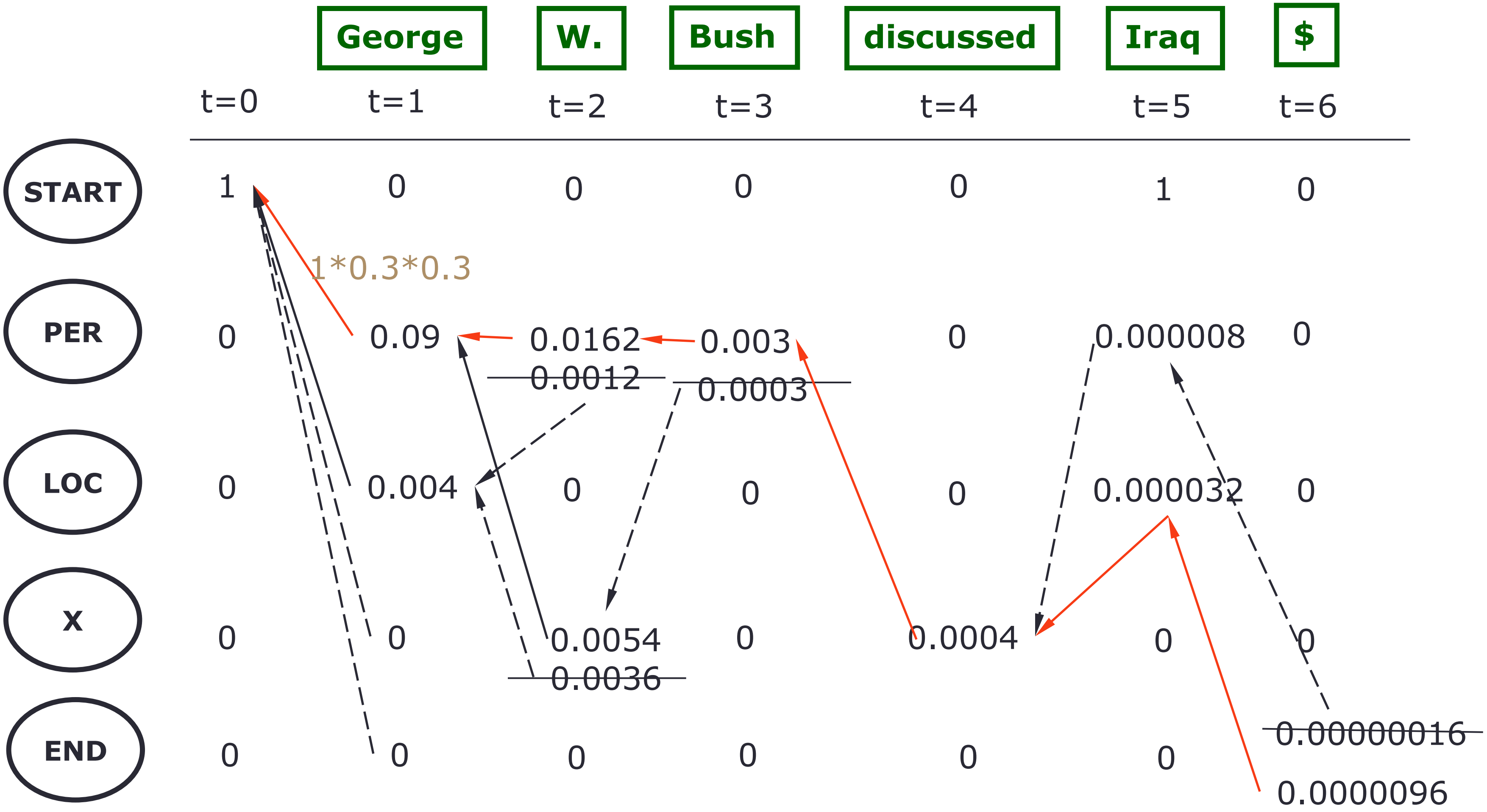
Slide from Heng Ji

Markov Chain for a Simple Name Tagger



Slide from Heng Ji

Viterbi Decoding of Name Tagger



	George	W.	Bush	discussed	Iraq	$
	t=0	t=1	t=2	t=3	t=4	t=5	t=6

	t=0	George t=1	W. t=2	Bush t=3	discussed t=4	Iraq t=5	$ t=6
START	1	0	0	0	0	1	0
PER	0	0.09	0.0162 ~~0.0012~~	0.003 ~~0.0003~~	0	0.000008	0
LOC	0	0.004	0	0	0	0.000032	0
X	0	0	0.0054 ~~0.0036~~	0	0.0004	0	0
END	0	0	0	0	0	0	~~0.00000016~~ 0.0000096

1*0.3*0.3

Current = Previous * Transition * Emission

Slide from Heng Ji

Limitations of HMMs

- Joint probability distribution $p(y, x)$
- Assume independent features
- Cannot represent overlapping features or long range dependencies between observed elements
  - Need to enumerate all possible observation sequences
  - Strict independence assumptions on the observations
- Toward discriminative/conditional models
  - Conditional probability P(label sequence y | observation sequence x) rather than joint probability P(y, x)
  - Allow arbitrary, non-independent features on the observation sequence X
  - The probability of a transition between labels may depend on past and future observations
  - Relax strong independence assumptions in generative models

Slide modified from Heng Ji

Maximum Entropy

- Why **maximum** entropy?
- Maximize entropy = Minimize commitment

- Model all that is known and assume nothing about what is unknown.
  - Model all that is known: satisfy a set of constraints that must hold

  - Assume nothing about what is unknown:
    choose the most “uniform” distribution
    ➔ choose the one with maximum entropy

Slide from Heng Ji

Why Try to be Uniform?

- Most Uniform = Maximum Entropy
- By making the distribution as uniform as possible, we don’t make any additional assumptions to what is supported by the data
- Abides by the principle of Occam’s Razor
  (least assumption = simplest explanation)
- Less generalization errors (less over-fitting)
  →more accurate predictions on test data

Slide from Heng Ji

Learning Coreference by Maximum Entropy Model

- Suppose that if the feature "Capitalization" = "Yes" for token t, then

  P (t is the beginning of a Name | (Captalization = Yes)) = 0.7

- How do we adjust the distribution?

  P (t is not the beginning of a name | (Capitalization = Yes)) = 0.3

- If we don't observe "Has Title = Yes" samples?

  P (t is the beginning of a name | (Has Title = Yes)) = 0.5

  P (t is not the beginning of a name | (Has Title = Yes)) = 0.5

Slide from Heng Ji

The basic idea

- Goal: estimate p

- Choose p with maximum entropy (or “uncertainty”) subject to the constraints (or “evidence”).

$$H(p) = -\sum_{x \in A \times B} p(x) \log p(x)$$

$$x = (a, b), \quad where \quad a \in A \wedge b \in B$$

Slide from Heng Ji

Setting

- From training data, collect (a, b) pairs:
  - a: thing to be predicted (e.g., a class in a classification problem)
  - b: the context
  - Ex: Name tagging:
    - a=person
    - b=the words in a window and previous two tags

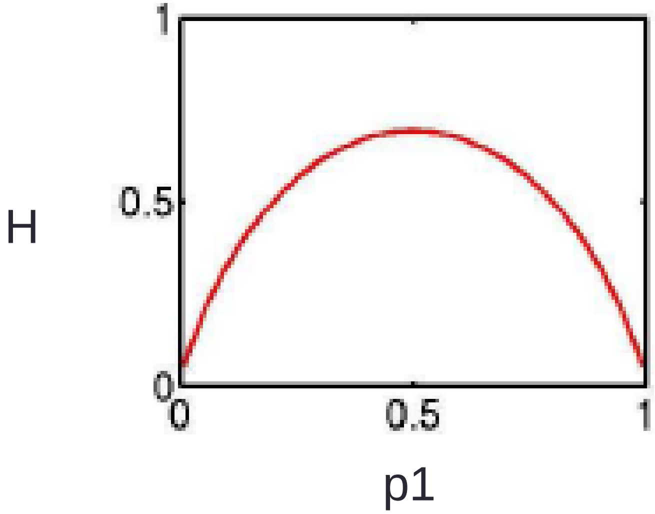
- Learn the prob of each (a, b): p(a, b)

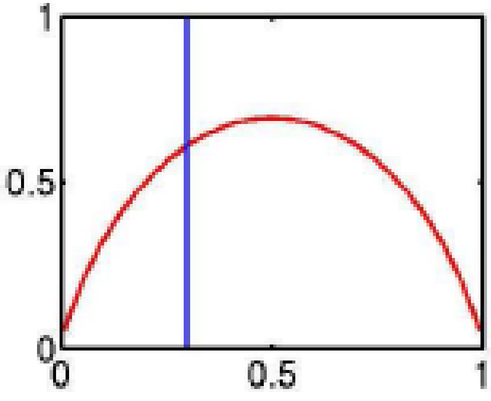
Slide from Heng Ji

Ex1: Coin-flip example (Klein & Manning 2003)

- Toss a coin: p(H)=p1, p(T)=p2.
- Constraint: p1 + p2 = 1
- Question: what’s your estimation of p=(p1, p2)?
- Answer: choose the p that maximizes H(p)

$$H(p) = -\sum_{x} p(x) \log p(x)$$

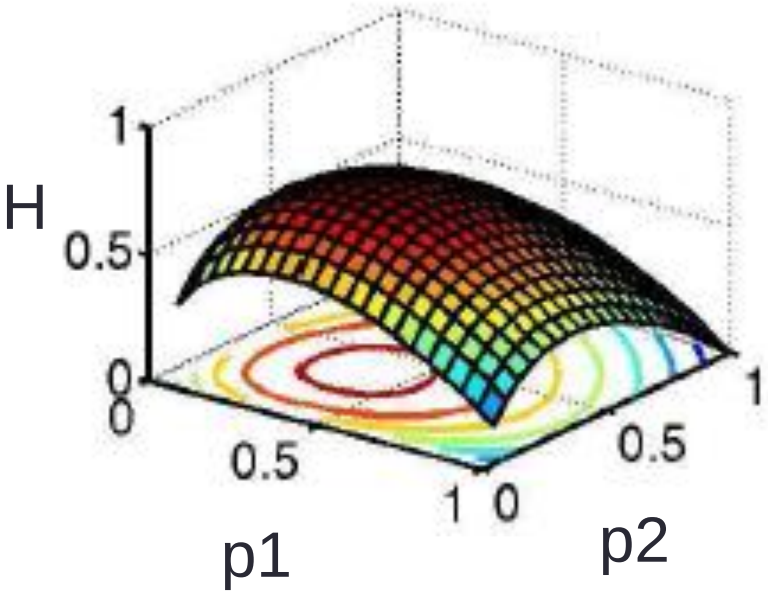


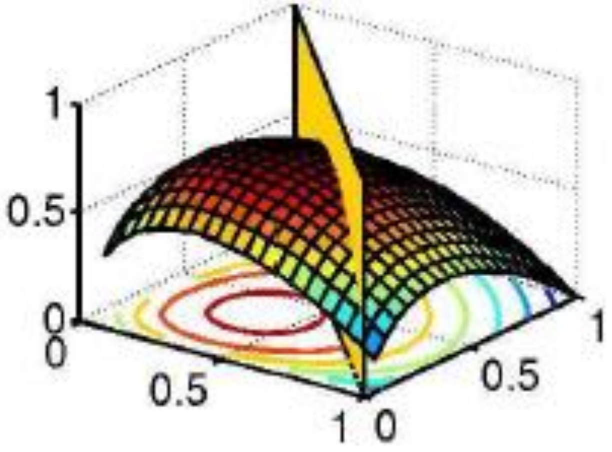


p1=0.3

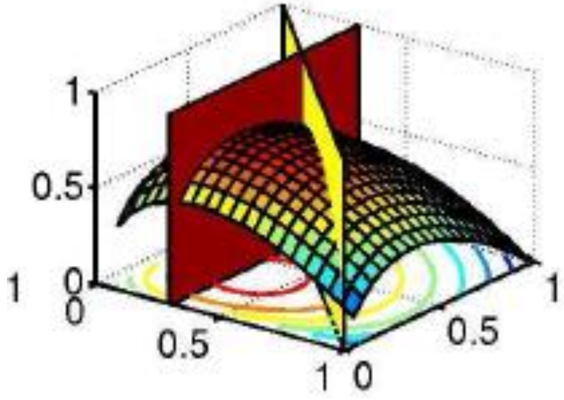
Slide from Heng Ji

Coin-flip example (cont)





p1 + p2 = 1



p1+p2=1.0, p1=0.3

Slide from Heng Ji

Ex2: An MT example (Berger et. al., 1996)

Possible translation for the word “in” is:

$$\{dans,\ en,\ \grave{a},\ au\ cours\ de,\ pendant\}$$

Constraint:

$$p(dans) + p(en) + p(\grave{a}) + p(au\ cours\ de) + p(pendant) = 1$$

Intuitive answer:

$$\begin{aligned} p(dans) &= 1/5 \\ p(en) &= 1/5 \\ p(\grave{a}) &= 1/5 \\ p(au\ cours\ de) &= 1/5 \\ p(pendant) &= 1/5 \end{aligned}$$

Slide from Heng Ji

An MT example (cont)

Constraints:

$$\begin{aligned} p(dans) + p(en) &= 3/10 \\ p(dans) + p(en) + p(\grave{a}) + p(au\ cours\ de) + p(pendant) &= 1 \end{aligned}$$

Intuitive answer:

$$\begin{aligned} p(dans) &= 3/20 \\ p(en) &= 3/20 \\ p(\grave{a}) &= 7/30 \\ p(au\ cours\ de) &= 7/30 \\ p(pendant) &= 7/30 \end{aligned}$$

Slide from Heng Ji

Why ME?

- Advantages
  - Combine multiple knowledge sources
    - Local
      - Word prefix, suffix, capitalization (POS - *(Ratnaparkhi, 1996)*)
      - Word POS, POS class, suffix (WSD - *(Chao & Dyer, 2002)*)
      - Token prefix, suffix, capitalization, abbreviation (Sentence Boundary - *(Reynar & Ratnaparkhi, 1997)*)
    - Global
      - N-grams *(Rosenfeld, 1997)*
      - Word window
      - Document title *(Pakhomov, 2002)*
      - Structurally related words *(Chao & Dyer, 2002)*
      - Sentence length, conventional lexicon *(Och & Ney, 2002)*
  - Combine *dependent* knowledge sources

Slide from Heng Ji

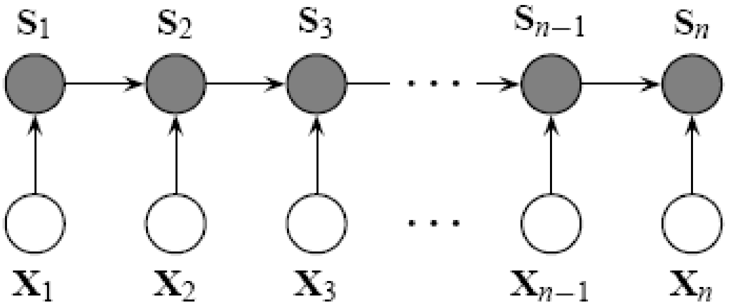
Why ME?

- Advantages
  - Add additional knowledge sources
  - Implicit smoothing
- Disadvantages
  - Computational
    - Expected value at each iteration
    - Normalizing constant
  - Overfitting
    - Feature selection
      - Cutoffs
      - Basic Feature Selection *(Berger et al., 1996)*

Slide from Heng Ji

Maximum Entropy Markov Models (MEMMs)

- A conditional model that representing the probability of reaching a state given an observation and the previous state
- Consider observation sequences to be events to be conditioned upon.



$$p(s \mid x) = p(s_1 \mid x_1) \prod_{i=2}^{n} p(s_i \mid s_{i-1}, x_i)$$

- Have all the advantages of Conditional Models
- No longer assume that features are independent
- Do not take future observations into account (no forward-backward)
- Subject to Label Bias Problem: Bias toward states with fewer outgoing transitions

Slide from Heng Ji

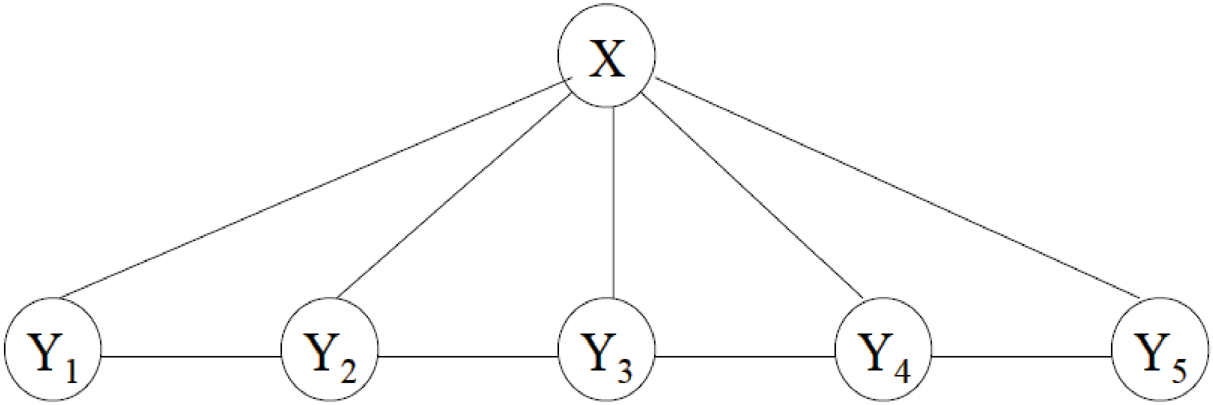
Conditional Random Fields (CRFs)

- Conceptual Overview
  - Each attribute of the data fits into a *feature function* that associates the attribute and a possible label
    - A positive value if the attribute appears in the data
    - A zero value if the attribute is not in the data
  - Each feature function carries a *weight* that gives the strength of that feature function for the proposed label
    - High positive weights: a good association between the feature and the proposed label
    - High negative weights: a negative association between the feature and the proposed label
    - Weights close to zero: the feature has little or no impact on the identity of the label
- CRFs have all the advantages of MEMMs without label bias problem
  - MEMM uses per-state exponential model for the conditional probabilities of next states given the current state
  - CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence
- Weights of different features at different states can be traded off against each other
- CRFs provide the benefits of discriminative models

Slide from Heng Ji

Example of CRFs

Suppose $P(Y_v \mid X, \text{all other } Y) = P(Y_v \mid X, \text{neighbors}(Y_v))$ then X with Y is a **conditional** random field



- $P(Y_3 \mid X, \text{all other } Y) = P(Y_3 \mid X, Y_2, Y_4)$
- Think of X as observations and Y as labels

Slide from Heng Ji

Sequential Model Trade-offs

	Speed	**Discriminative vs. Generative**	**Normalization**
HMM	very fast	generative	local
MEMM	mid-range	discriminative	local
CRF	relatively slow	discriminative	global

Slide from Heng Ji

State-of-the-art and Remaining Challenges

- State-of-the-art Performance
  - On ACE data sets: about 89% F-measure (Florian et al., 2006; Ji and Grishman, 2006; Nguyen et al., 2010; Zitouni and Florian, 2008)
  - On CONLL data sets: about 91% F-measure (Lin and Wu, 2009; Ratinov and Roth, 2009)

- Remaining Challenges
  - Identification, especially on organizations
    - Boundary: “Asian Pulp and Paper Joint Stock Company , Lt. of Singapore”
    - Need coreference resolution or context event features: “**FAW** has also utilized the capital market to directly finance, and now *owns* three domestic listed *companies*” (*FAW = First Automotive Works*)
  - Classification
    - “Caribbean Union”: ORG or GPE?

Slide from Heng Ji

Slides

- The slides on machine learning are from **Heng Ji**, who is a IE researcher at RPI
- Literature:
  - Dan Klein and Chris Manning. Maxent Models, Conditional Estimation, and Optimization, without the Magic. Tutorial presented at NAACL 2003 and ACL 2003.
    - Available from Dan Klein's web page (at the bottom):
    - http://www.cs.berkeley.edu/~klein
  - See also the two papers mentioned in the slides:
    - Ratnaparkhi's 1998 thesis
    - Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics* (22-1). March 1996
  - CRF (and MEMM) paper:
    - John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" *Departmental Papers (CIS)* (2001). Available at: http://works.bepress.com/andrew_mccallum/4

- Thank you for your attention!