Named Entity Recognition and Classification

Asif Ekbal
AI-NLP-ML Group
Dept. of Computer Science and Engineering
IIT Patna, India-800 013
Email: asif@iitp.ac.in
asif.ekbal@gmail.com

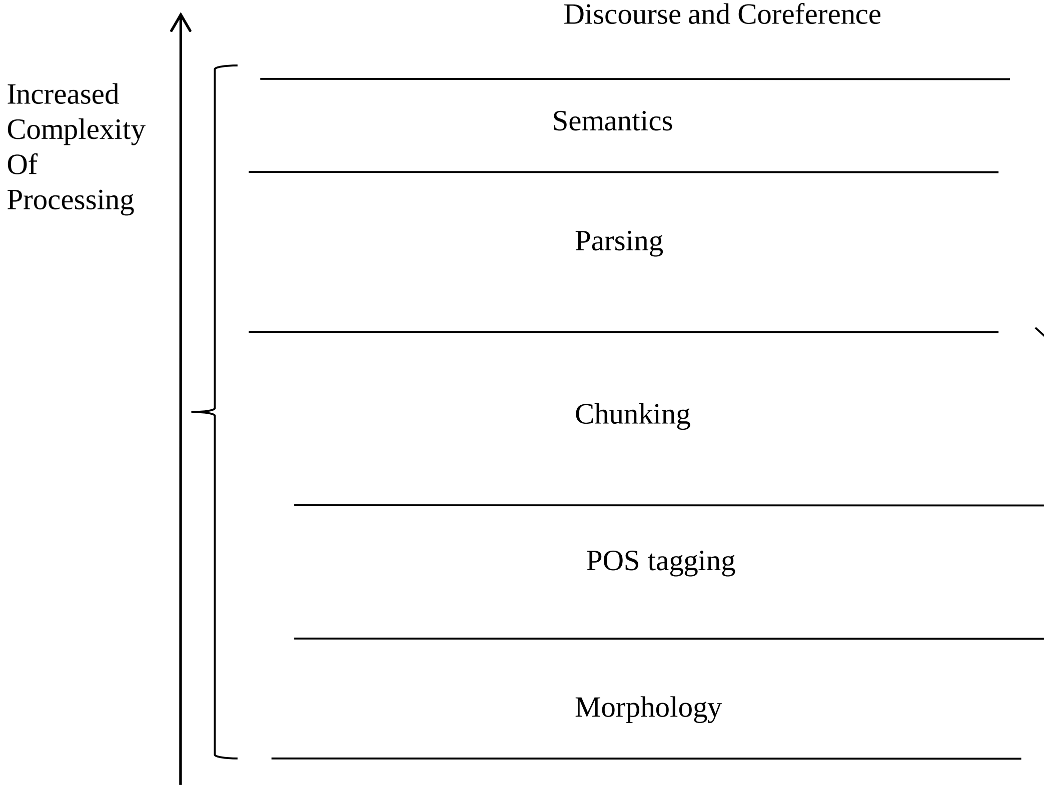
Deep Learning Techniques for Conversational AI
April 15, 2022

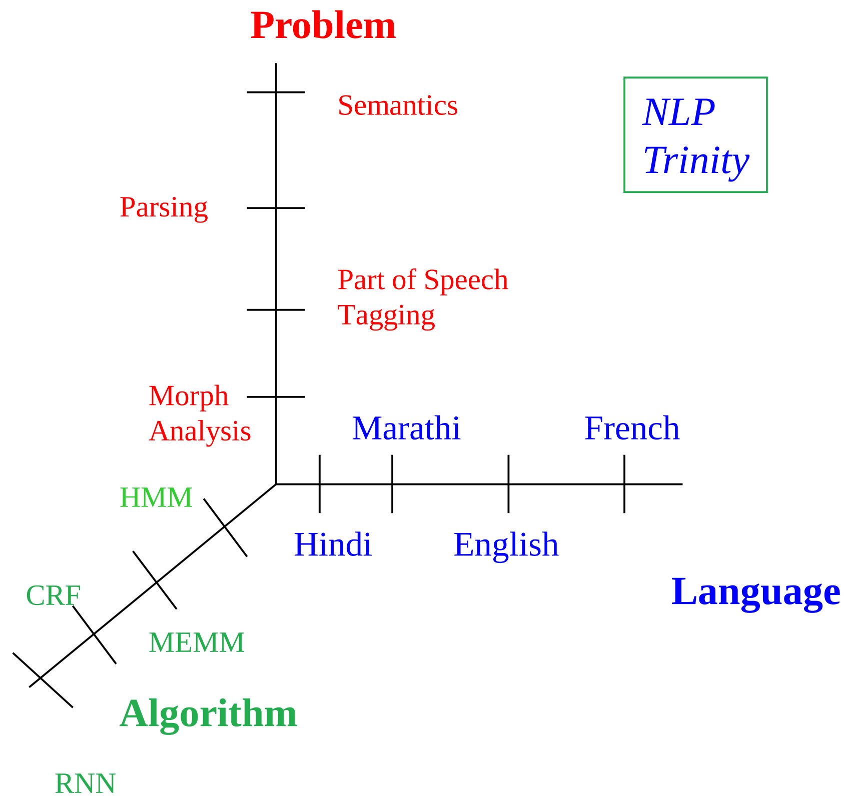
Outline

- Background: *NLP and its impact*
- Introduction to the various issues of NER
- NER in different languages
- NER in Indian languages

- Weighted Vote based Classifier Ensemble
  - Introduction to GA
  - Some Issues of Classifier Ensemble

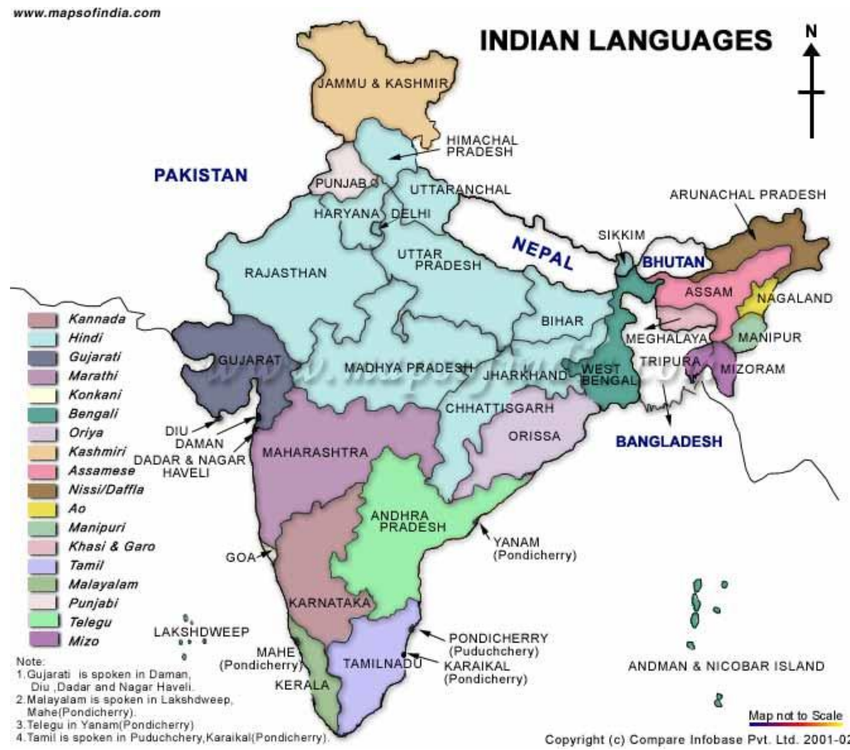
NLP Trinity


Discourse and Coreference
Semantics
Parsing
Chunking
POS tagging
Morphology
Increased Complexity Of Processing


Problem
Semantics
Parsing
Part of Speech Tagging
Morph Analysis
NLP Trinity
Marathi
French
Hindi
English
Language
HMM
CRF
MEMM
Algorithm
RNN

Multilinguality: Indian situation

- Major streams
  - Indo European
  - Dravidian
  - Sino Tibetan
  - Austro-Asiatic
- Some languages are ranked within 20 in in the world in terms of the populations speaking them
  - Hindi : 4th (~350 milion)
  - Bangla: 5^{h} (~230 million)
  - Marathi 10th (~84 million)



Language Technology or Natural Language Processing: Background & Relevance in Indian Scenario

Background: Indian Context

- India is a multi-lingual country with great linguistic and cultural diversities
- 22 official languages mentioned in the Indian constitution
- However, Census of India in 2001 reported-
  - **122 major languages**
  - **1,599 other regional languages**
  - **13 scripts**
  - **720 dialects**
  - **30 languages** are spoken by more than **one million native speakers**
  - **122** are spoken by more than **10,000 people**
- **20%** understand English
- **80%** cannot understand

Background

- Phenomenal growth in the number of internet users, social media (***Facebook, Twitter*** etc.)
- Increasing tendency of using Indian language contents for exchanging information
- **Digital divide** cannot be tackled unless citizens are given flexibility in **communicating in their own languages**

Language Technology or Natural Language Processing (NLP) that deals with developing theories and techniques for effective communication in human languages play an important role towards creating this digital society

TDIL: MeiTY, Govt. of India

- Technology Development for Indian Languages (TDIL) Programme
- **Objective**:
  - developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier;
  - creating and accessing multilingual knowledge resources; and
  - integrating them to develop innovative user products and services

TDIL: Some major initiatives

- Development of English to Indian Language Machine Translation (**Anuvadaksh**):

English to Hindi/Marathi/Bangla/Oriya/Tamil/Urdu/Gujrati/Bodo

- Development of English to Indian Language Machine Translation System with Angla-Bharti Technology: English to Bangla/Punjabi/Malaylam/Urdu/Hindi/Telugu
- Development of Indian Language to Indian Language Machine Translation System (**Sampark**)- 18 pairs of languages

-Hindi to Bengali, Bengali to Hindi, Marathi to Hindi, Hindi to Marathi, Hindi to Punjabi, Punjabi to Hindi, Hindi to Tamil, Tamil to Hindi, Hindi to Kannada, Kannada to Hindi, Hindi to Telugu, Telugu to Hindi, Hindi to Urdu, Urdu-Hindi, Malaylam to Tamil, Tamil to Malaylam, Tamil to Telugu, Telugu to Tamil

TDIL: **Major initiatives**

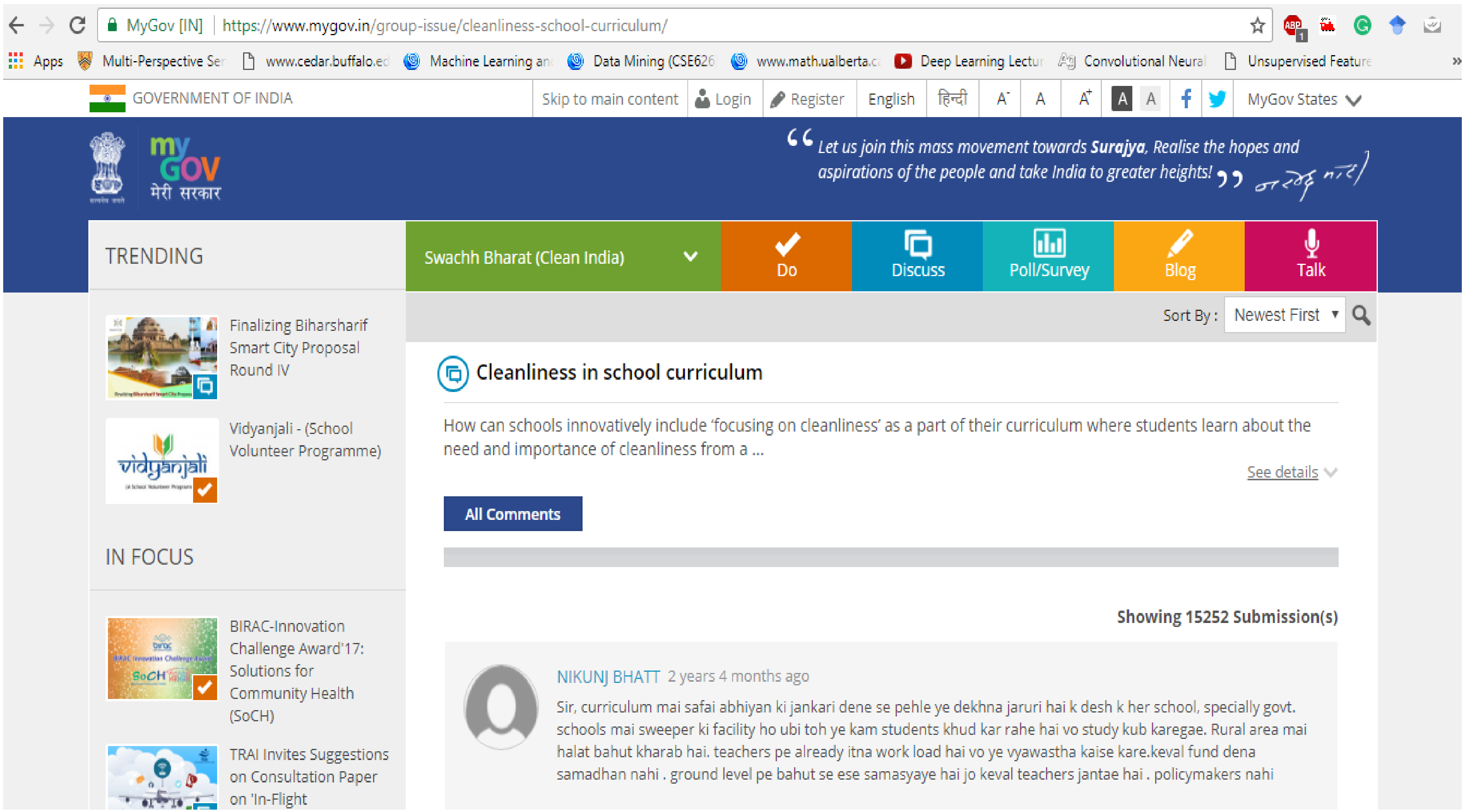
- Development of Cross-Lingual Information Access (CLIA)
  - Assamese, Bengali, Hindi, Oriya, Punjabi, Tamil, Telugu, Marathi

- Development of Robust Document Analysis & Recognition System for Indian Languages (OCR)-14 languages
  - Assamese, Bengali, Devanagri, Gujrati, Gurumukhi, Kannada, Malaylam, Manipuri, Marathi, Oriya, Tamil, Telugu, Tibetan, Urdu

- Development of Text to Speech System in Indian Languages
- Development of Automatic Speech Recognition System in Indian Languages
- *Development of Hindi to English Machine Translation in Judicial Domain*

A Case-Study: **MyGov.in Portal**

Govt. Portal: MyGov.in



Govt. Portal: **MyGov.in**

- **Citizen-centric platform** empowers people to connect with the Government & contribute towards good governance
- Unique first of its kind participatory governance initiative involving the common citizen at large
- Idea is to bring the government closer to the common man by the **use of online platform** creating an interface for **healthy exchange of ideas** and **views** involving the common citizen and experts
- Ultimate goal is to contribute to the **social and economic transformation of India**
- Was launched on July 26, 2014 by the Hon'ble PM

Govt. Portal: **MyGov.in**

- This has been more than successful in keeping the citizens engaged on important policy issues and governance, be it **Clean Ganga**, **Girl Child Education**, **Skill Development** and **Healthy India** to name a few
- Has become a key part of the **policy and decision making** process of the country

- Platform has been able
  - to provide the citizens a voice in the governance process of the country and
  - create grounds for the citizens to become stakeholders not only in policy formulation and recommendation but also implementation through actionable tasks

Govt. Portal: **MyGov.in**

- **Major attributes**: Discussion, Tasks, Talks, Polls and Blogs on various groups based on the diverse governance and public policy issues

- Has more than **1.78 Million users** who contribute their ideas through discussions and also participate through the various earmarked tasks

- Platform gets more than **10,000 posts per weeks** on various issues

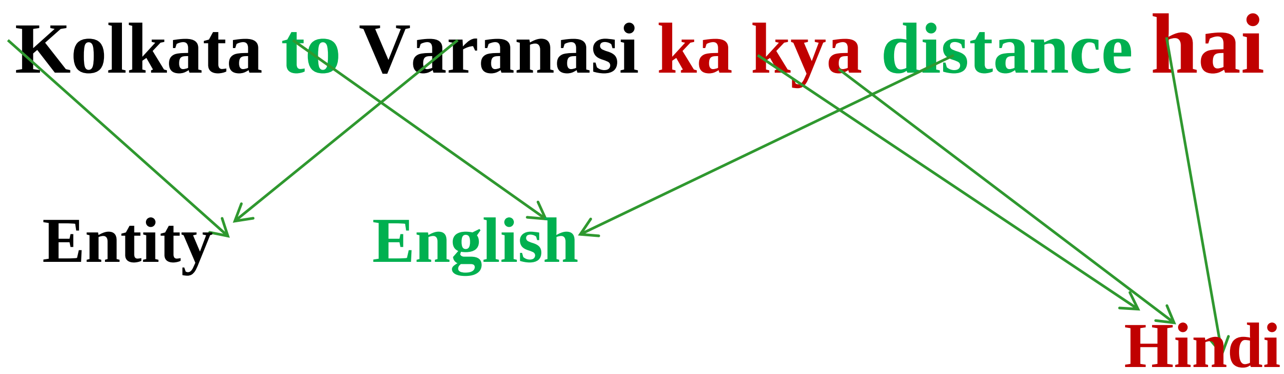
Feedbacks are analyzed and put together as suggestions for the concerned departments which are responsible to transform them into actionable agenda

- Infeasible to **mine** the most **relevant information** from this huge data

- Needs a method for automated analysis of this data
  - **Demands sophisticated NLP and ML techniques to build these**

Code-mixing

- Code-mixing refers to the mixing of two or more languages or language varieties in speech/text



Code-Mixing in MyGov.in: Few Examples

- *Sir ji aapka ye abhiyan acha ha isse naye bharat ka nirman hoga maine apne school ke student ke sath milkar hospital ki safai ki and jagrukta rali nikali jisse log gandagi kam failaye.*
- *Aaj her school main swachta abhiyan honi chye we do it*
- *india ko clean rakhne ke lie gandgi karne walo pe penalty lagani chahiye jo kaam das sal me hoga penalty lagane ke bad wo kuch hi dino me ho jaega*
- *Modi sir swachh bharat m aapke bjp poltician photo click krawane k liye safai krte h sathinye neta sirf pik click krte h bs.*
-
- *Our School also participated in Clean India Campaign . The students of class XII cleaned a Park and a Basket Ball area .*

Why to Analyse?

- Public opinions play important roles for the betterment of human lives
- Huge volumes and varieties of user-generated contents and user interaction networks constitute new opportunities for understanding social behavior
- Understanding deep feeling of public can help government to anticipate deep social changes and adapt to population expectations

Discipline known as Opinion Mining or Sentiment Analysis

NLP: Projected Growth

- *Growing in an exponential manner*
- *Expected to touch the market of* ***$16 billion in 2021***
    - *With compound growth rate of 16% annually*
- Reasons behind this growth
    - Rising of the Chatbots
    - Urge of discovering the customer insights
    - Transfer of technology of messaging from manual to automated
    - Translation of contents, and
    - many other tasks which are required to be automated and involve language/Speech at some point
    - *Etc.*

Major Industries: *Amazon, Google, Microsoft, Facebook, IBM etc.*

NLP: Evolution

- Evolving from ***human-computer interaction*** to ***human-computer conversation***

- The first critical part of NLP Advancements – Biometrics

- The second critical part of NLP advancements–Humanoid Robotics

NLP: In Governance

- NLP techniques for the delivery to the common people and to decrease the interaction gap between the citizen and the Government
- **Uses of NLP in Government Websites**
  - Making e-governance related information to be available in multiple languages
- **Natural Language Generation in e-Governance**
  - Chatbot
  - E.g. farmer can not read or write, but with the multilingual support and NLP generation, s/he can communicate the query in any language and get it resolved

NLP: In Business, Healthcare

- **Sentiment Analysis:** Analyzing public opinion
- **Email Filters:** Filtering out irrelevant emails
- **Voice Recognition:** Developing smart voice-driven services
- **Information Extraction**
- **NLP in Healthcare**
  - main concern and priority in nowadays the healthcare system is to provide better and 24/7 EHR experience
  - Voice-support systems, Predictive systems, Prescriptive analytics
- **NLP in Healthcare: Multilingual**
  - can be used to reduce the communication and interaction gap between Healthcare technologies (such as patient portals which contain health records of a patient) and patients
  - Patients can interact in his/her own language
  - Easier for a patient to understand health status

NLP: In Healthcare

- **Increasing the dimension of high quality of care**
  - Healthcare reports generally contain parameters which require proper attention
  - Use of NLP can provide significant relief in case of calculating the measure of inpatient care and monitoring the clinical guidelines

- **Identification of the patients which require Improved Care Coordination**
  - Automated detection of cancer, detection of the root causes related to any substance disorder are some of the examples

NLP: In Finance

- **Credit Scoring Method**
  - **Estimate risk factor of giving loan with the past histories**
  - **E.g.** Lenddo EFL (with 115 employees), a Singapore-based company developed a software called Lenddo Score which uses machine learning and NLP to assess and calculate an individual's creditworthiness.

- **Document search**
  - Nuance Communications based in Massachusetts developed software known as Nuance Document Finance Solution, which is used to aid financial services companies in automatizing the documentation process

- **Fraud detection in banking**
- **Stock market prediction-** based on sentiment

NLP: In Other domains

- **National Security**
  - Sentiment in Cross-border languages
  - Hate Speech, Radicalization

- **NLP in Recruitment**

  - Searching the appropriate applications from the data, and it also can be used for selecting the best applications from the data available

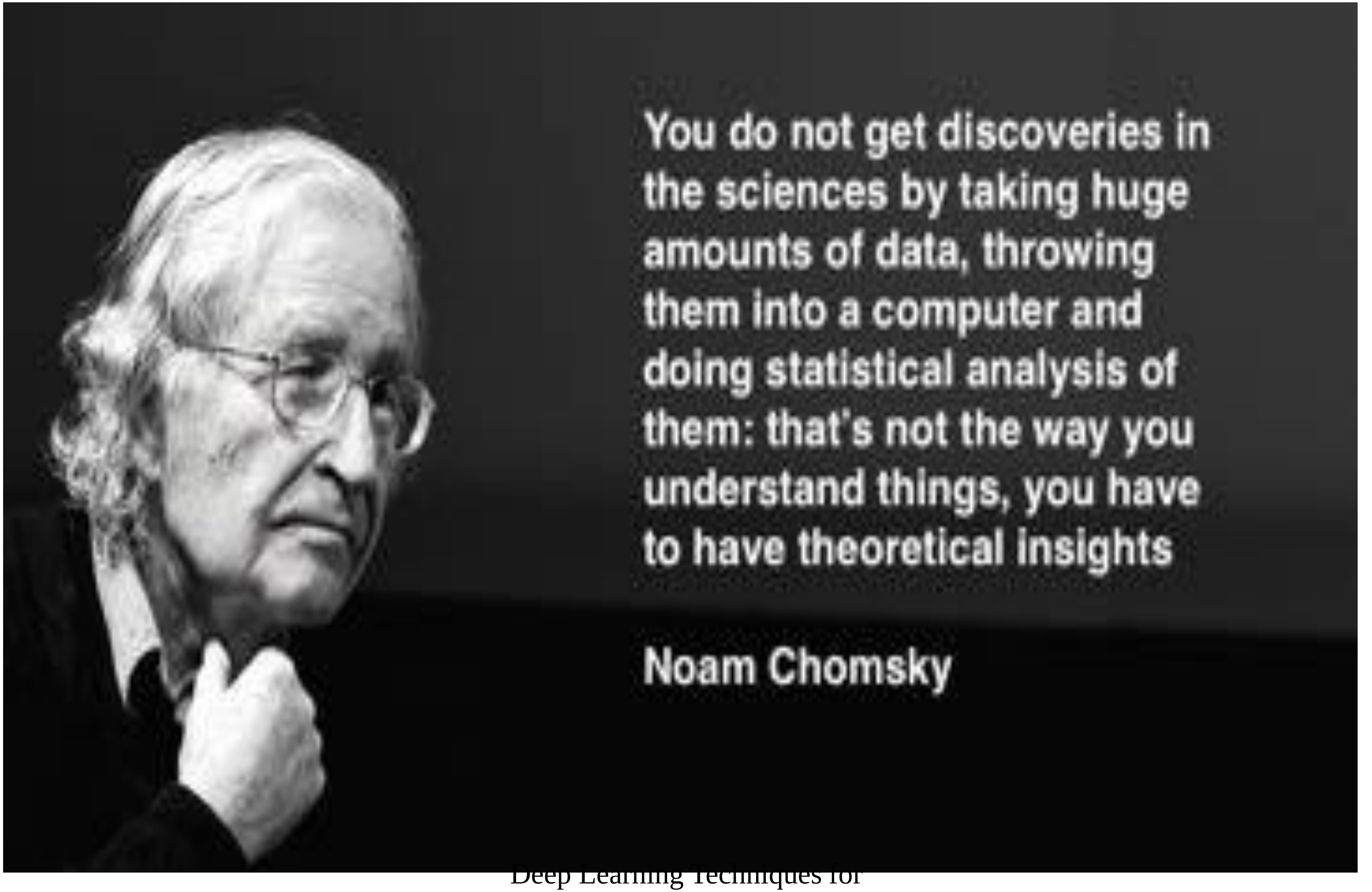
NLP and ML: *From Past to Present*

- **NLP based systems have enabled wide-range of applications**
  - Google's powerful search engines, Google's MT
  - Alexa etc.
  - Amazon Comprehend Medical services
  - Cognitive Analytics and NLP, Spam detection, NLP in Recruitment
  - Sentiment Analysis, Hate Speech detection, Fake News detection
- **Shallow ML algorithms (corresponds to Statistical NLP)**
  - Used extensively (HMM, MaxEnt, CRF, SVM, Logistic Regression etc.)
  - Requires handcrafting of features
  - Time-consuming
  - Curse of dimensionality (because of joint modeling of language models)

NLP and ML: From Past to Present

- Deep Learning algorithms
  - No feature engineering
  - Success of distributed representations (Neural language models)
- Some recent developments
  - The rise of distributed representations (e.g., Word2vec, GLOVE, ELMO, BERT etc)
  - Convolutional, recurrent, recursive neural networks, Transformer, Reinforcement learning
  - Unsupervised sentence representation learning
  - Combining deep learning models with memory-augmenting strategies
- Explainable AI

Statistics are no panacea!



Background

Background: Information Extraction

- To extract information that fits pre-defined database schemas or templates, specifying the output formats

- **IE Definition**
  - **Entity**: an object of interest such as a person or organization
  - **Attribute**: A property of an entity such as name, alias, descriptor or type
  - **Fact**: A relationship held between two or more entities such as Position of Person in Company
  - **Event**: An activity involving several entities such as terrorist act, airline crash, product information

The Problem

Date

Time: Start - End

DATE: Friday, March 24, 2006
TIME: 9:30-11:00 a.m.
LOCATION: 1014 DOW

Location

SPEAKER: Dave Lewis

Speaker

TITLE: Bayesian Logistic Reg ssification and Mining (Plus A Big New Test Collection)

ABSTRACT
Bayesian logistic regression allows incorporating task knowledge through model structure and priors on parameters. I will discuss content-based text categorization and authorship attribution using 1) priors that control sparsity and sign of parameters, 2) priors that incorporate domain knowledge from reference books and other texts, and 3) the use of polytomous (1-of-k) dependent variables. All experiments were performed with our open-source programs, BBR and BMR, which can fit models with millions of parameters. (Joint work with David Madigan, Alex Genkin, Aynur Dayanik, Dmitriy Fradkin, and Vladimir Menkov at Rutgers and DIMACS.) I will also briefly discuss the IIT CDIP (Complex Document Information Processing) test collection, which I am developing under an ARDA subcontract to Illinois Institute of Technology. It is based on 1.5TB of scanned and OCR'd documents released in tobacco litigation, and will be a major resource for research in information retrieval, document analysis, social network analysis, and perhaps databases. (Joint work with Gady Agam, Shlomo Argamon, Ophir Frieder, Dave Grossman, reds.)

Person

BIOGRAPHY
Dave Lewis is based in Chicago, IL, and consults on information retrieval, data mining, and natural language processing. He previously held research positions at AT&T Labs, Bell Labs, and the University of Chicago. He received his Ph.D. in Computer Science from the University of Massachusetts, Amherst, and did his undergraduate work down the road at Michigan State.

What is "Information Extraction"?

As a task: **Filling slots in a database from sub-segments of text.**

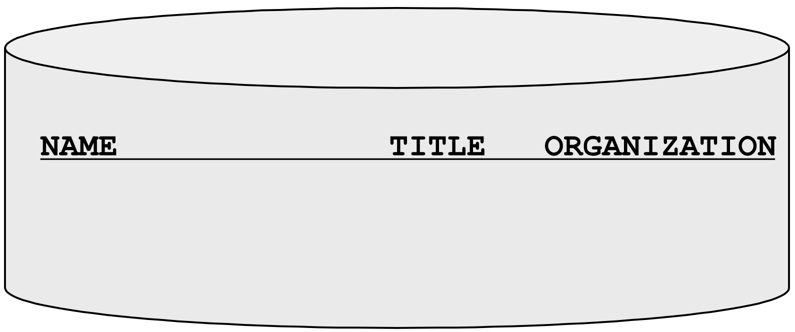
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



Courtesy of William W. Cohen

What is "Information Extraction"?

As a task: **Filling slots in a database from sub-segments of text.**

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Courtesy of William W. Cohen

What is "Information Extraction"?

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka "named entity extraction"

Courtesy of William W. Cohen

What is "Information Extraction"?

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Courtesy of William W. Cohen

What is “Information Extraction”?

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation CEO Bill Gates
Microsoft Gates

Microsoft

Bill Veghte Microsoft VP
Richard Stallman founder Free Software Foundation

Courtesy of William W. Cohen

What is "Information Extraction"?

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

* Microsoft
Bill Veghte

* Microsoft
VP

Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Courtesy of William W. Cohen

What is Named Entity Recognition and Classification (NERC)?

- NERC – Named Entity Recognition and Classification (NERC) involves identification of proper names in texts, and classification into a set of pre-defined categories of interest as:
  - Person names (names of people)
  - Organization names (companies, government organizations, committees, etc.)
  - Location names (cities, countries etc)
  - Miscellaneous names (Date, time, number, percentage, monetary expressions, number expressions and measurement expressions)

Named Entity Recognition

Markables (as defined in MUC6 and MUC7)

- Names of **organization**, **person**, **location**
- Mentions of **date** and **time**, **money** and **percentage**

Example:

"Ms. **Washington**'s candidacy is being championed by several powerful lawmakers including her boss, Chairman **John Dingell** (D., **Mich**.) of the **House Energy and Commerce Committee**."

Task Definition

- Other common types: measures (percent, money, weight etc), email addresses, web addresses, street addresses, etc.
- Some domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.
- MUC-7 entity definition guidelines (Chinchor'97)

http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html

Basic Problems in NER

- Generative in nature
- Variation of NEs – e.g. Prof Manning, Chris Manning, Dr Chris Manning
- Ambiguity of NE types:
  - Washington (location vs. person)
  - May (person vs. month)
  - Ford (person vs organization)
  - 1945 (date vs. time)
- Ambiguity with common words, e.g. "*Kabita*"

  - Name of person vs. poem

More complex problems in NER

- Issues of style, structure, domain, genre etc.
- Punctuation, spelling, spacing, formatting, … all have an impact:

Dept. of Computing and Maths

Manchester Metropolitan University

Manchester

United Kingdom

Applications

- Intelligent document access
  - Browse document collections by the entities that occur in them
  - Application domains:
    - News
    - Scientific articles, e.g, MEDLINE abstracts
- Information retrieval and extraction
  - Augmenting a query given to a retrieval system with NE information, more refined information extraction is possible
  - For example, if a person wants to search for document containing '*kabiTA*' as a proper noun, adding the NE information will eliminate irrelevant documents with only '*kabiTA*' as a common noun

Applications

- Machine translation
  - NER plays an important role in translating documents from one language to other
  - Often the NEs are transliterated rather than translated
  - For example, '*yAdabpur bishvabidyAlaYa*' → '*Jadavpur University*'
- Automatic Summarization
  - NEs given more priorities in deciding the summary of a text
  - Paragraphs containing more NEs are most likely to be included into the summary

Applications

- Question-Answering Systems
  - NEs are important to retrieve the answers of particular questions

- Speech Related Tasks
  - In Text to Speech (TTS), NER is important for identifying the number format, telephone number and date format
  - In speech rhythm- necessary to provide a short break after the name of person
  - Solving **Out Of Vocabulary (OOV)** words is important in speech recognition

Corpora, Annotation

Some NE Annotated Corpora

- MUC-6 and MUC-7 corpora - English
- CONLL shared task corpora
  - http://cnts.uia.ac.be/conll2003/ner/ : NEs in English and German
  - http://cnts.uia.ac.be/conll2002/ner/ : NEs in Spanish and Dutch
- ACE – English - http://www.ldc.upenn.edu/Projects/ACE/
- TIDES surprise language exercise (NEs in Hindi)
- NERSSEAL shared task- NEs in Bengali, Hindi, Telugu, Oriya and Urdu (http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5)

Corpora, Annotation

- Biomedical, Biochemical and Health Corpora
  - BioNLP-04 shared task
  - BioCreative shared tasks
  - AiMed
  - I2B2
- NER in Tweet
  - ACL-IJCNLP Workshop on Noisy User-generated Text (W-NUT)

The MUC-7 Corpus

<ENAMEX TYPE="LOCATION">CAPE CANAVERAL</ENAMEX>, <ENAMEX TYPE="LOCATION">Fla.</ENAMEX> &MD; Working in chilly temperatures <TIMEX TYPE="DATE">Wednesday</TIMEX> <TIMEX TYPE="TIME">night</TIMEX>, <ENAMEX TYPE="ORGANIZATION">NASA</ENAMEX> ground crews readied the space shuttle Endeavour for launch on a Japanese satellite retrieval mission.

<p>

Endeavour, with an international crew of six, was set to blast off from the <ENAMEX TYPE="ORGANIZATION|LOCATION">Kennedy Space Center</ENAMEX> on <TIMEX TYPE="DATE">Thursday</TIMEX> at <TIMEX TYPE="TIME">4:18 a.m. EST</TIMEX>, the start of a 49-minute launching period. The <TIMEX TYPE="DATE">nine day</TIMEX> shuttle flight was to be the 12th launched in darkness.

Performance Evaluation

- Evaluation metric – mathematically defines how to measure the system's performance against a human-annotated, gold standard

- Scoring program–implements the metric and provides performance measures
  - For each document and over the entire corpus
  - For each type of NE

The Evaluation Metric

Precision = correct answers/answers produced

Recall = correct answers/total possible correct answers

Trade-off between precision and recall

F-Measure = $(\beta^2 + 1)PR / \beta^2 R + P$

β reflects the weighting between precision and recall, typically $\beta=1$

The Evaluation Metric (2)

$$\text{Precision} = \frac{\text{Correct} + \tfrac{1}{2}\,\text{Partially correct}}{\text{Correct} + \text{Incorrect} + \text{Partial}}$$

$$\text{Recall} = \frac{\text{Correct} + \tfrac{1}{2}\,\text{Partially correct}}{\text{Correct} + \text{Missing} + \text{Partial}}$$

NE boundaries are often misplaced, so some partially correct results

Named Entity Recognition

- Handcrafted systems
  - Knowledge (rule) based
    - Patterns
    - Gazetteers
- Automatic systems
  - Statistical
  - Machine learning-*Supervised*, *Semi-supervised*, *Unsupervised*
- Hybrid systems

Comparisons between two Approaches

Knowledge Engineering

- rule based
- developed by experienced language engineers
- makes use of human intuition
- requires only small amount of training data
- development could be very time consuming
- some changes may be hard to accommodate

Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- requires large amounts of annotated training data
- annotators are cheap (but you get what you pay for!)
- easily trainable and adaptable to new domains and languages

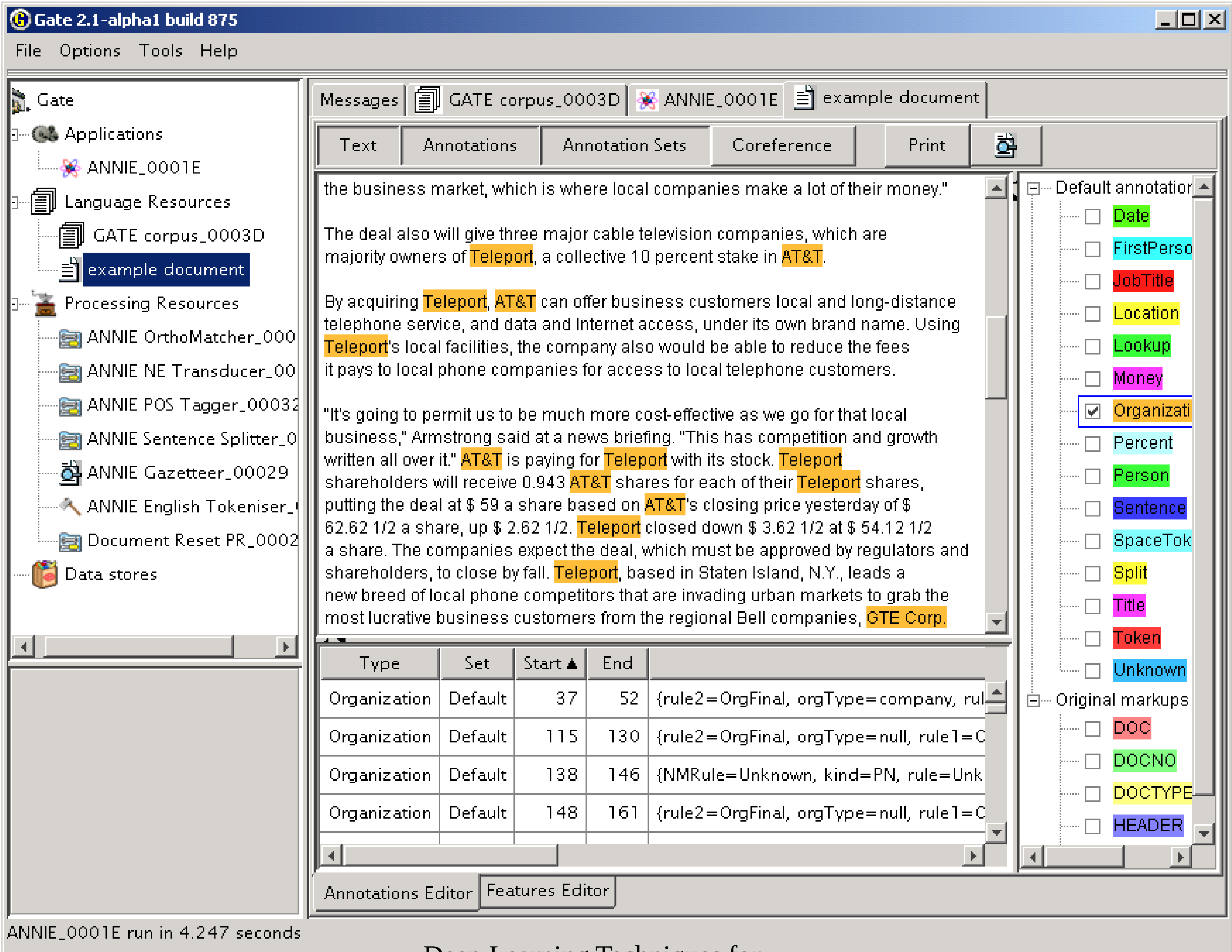
Named Entity Recognition

- Handcrafted systems
  - LTG (Mikheev et al., 1997)
    - 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

Named Entity Recognition

- 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)
  - FACILE (1998)- Fast and Accurate Categorisation of Information by Language Engineering
    - NEA language (Named Entity Analysis)
    - Context-sensitive rules
  - NetOwl (MUC-7)
    - Commercial product
    - C++ engine, extraction rules

Named Entities in GATE



NER–automatic approaches

- Learning of statistical models or symbolic rules
  - Use of annotated text corpus
    - Manually annotated
    - Automatically annotated

- ML approaches frequently break down the NE task in two parts:
  - Recognising the entity boundaries
  - Classifying the entities in the NE categories

NER – automatic approaches

- Tokens in text are often coded with the IOB scheme
  - O – outside, B-XXX – first word in NE, I-XXX – all other words in NE

  e.g.

India	B-LOC
played	O
with	O
Vivian	B-PER
Richards	I-PER

  - Probabilities:
    - Simple:
      - P(tag i | token i)
    - With external evidence:
      - P(tag i | token i-1, token i, token i+1)

NER–automatic approaches

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

NER – automatic approaches

- HMM-*Generative model*
  - Markov models, Viterbi
  - Works well when large amount of data is available: Nymble (1997) / IdentiFinder (1999)

- Maximum Entropy (ME)-*Discriminative model*
  - 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%

ML features

- The choice of features
  - Lexical features (words)
  - Part-of-speech
  - Orthographic information
  - Affixes (prefix and suffix of any word)
  - Gazetteers

- External, unmarked data is useful to derive gazetteers and for extracting training instances

IdentiFinder [Bikel et al 99]

- Based on Hidden Markov Models
- 7 regions of HMM–one for each *MUC type*, *not-name*, *begin-sentence* and *end-sentence*

- Features
  - Capitalisation
  - Numeric symbols
  - Punctuation marks
  - Position in the sentence
  - 14 features in total, combining above info, e.g., containsDigitAndDash (09-96), containsDigitAndComma (23,000.00)

IdentiFinder (2)

- Evaluation: MUC-6 (English) and MET-1(Spanish) corpora

- Mixed case English
  - IdentiFinder - 94.9% F-measure
  - Best rule-based – 96.4% F-measure
- Spanish mixed case
  - IdentiFinder – 90% F-measure
  - Best rule-based - 93% F-measure
  - Lower case names, noisy training data, less training data

- Impact of size of data- Trained with 650,000 words, but similar performance with half of the data. Less than 100,000 words reduce the performance to below 90% on English

MENE [Borthwick et al 98]

- Rule-based NE + ML based NE- achieve better performance

- Tokens tagged as: XXX_start, XXX_continue, XXX_end, XXX_unique, other (non-NE), where XXX is an NE category

- Uses Maximum Entropy (ME)
  - One only needs to find the best features for the problem
  - ME estimation routine finds the best relative weights for the features

MENE (2)

- Features
  - Binary features–“token begins with capitalised letter”, “token is a four-digit number”
  - Lexical features–dependencies on the surrounding tokens (window ±2) e.g., “Mr” for people, “to” for locations
  - Dictionary features–equivalent to gazetteers (first names, company names, dates, abbreviations)
  - External systems–whether the current token is recognised as a NE by a rule-based system

MENE (3)

- MUC-7 formal run corpus
  - MENE – *84.2%* F-measure
  - Rule-based systems– *86% - 91 %* F-measure
  - MENE + rule-based systems – *92%* F-measure

- Learning curve
  - 20 docs – 80.97% F-measure
  - 40 docs – 84.14% F-measure
  - 100 docs – 89.17% F-measure
  - 425 docs – 92.94% F-measure

Named Entity Recognition: Maximum Entropy Approach Using Global Information

(*Chieu and Ng, 2003*)

Global Information

- Local Context is insufficient
  - "**Mary Kay** Names Vice Chairman…"

- Global Information is useful
  - "Richard C. Bartlett was named to the newly created position of vice chairman of **Mary Kay Corp**."

Named Entity Recognition

- Modeled as a classification problem

- Each token is assigned one of 29 (= 7*4 + 1) classes:
  - person_begin, person_continue, person_end, person_unique
  - org_begin, org_continue, org_end, org_unique,
  - …
  - nn (not-a-name)

Named Entity Recognition

Consuela Washington , a longtime

person_begin person_end nn nn nn

House staffer ... the Securities and

org_unique nn nn org_begin org_continue

Exchange Commission in the Clinton …

org_continue org_end nn nn person_unique

Maximum Entropy Modeling

The distribution *p** in the conditional ME framework:

$$p^*(s_i \mid s_{i-1}, o) = \frac{1}{Z(s_{i-1}, o)} \sum_a \exp(\alpha_a f_a(s_i, o))$$

$f_j(h,o)$: binary feature
α_j : parameter / weight of each feature

Java-based opennlp maxent package:
http://maxent.sourceforge.net

Checking for Valid Sequence

- To discard invalid sequences like:
  - person_begin location_end …
- Transition probability $P(c_i \mid c_{i-1}) = 1$ if a valid transition, 0 otherwise
  - Dynamic programming to determine the valid sequence of classes with highest probability

$$P(c_1, \ldots, c_n | s, D) = \prod_{i=1}^{n} P(c_i | s, D) * P(c_i | c_{i-1})$$

Local Features

- Case and zone
  - initCaps, allCaps, mixedCaps
  - TXT, HL, DATELINE, DD
- First word
- Word string
- Out-of-vocabulary
  - WordNet

Local Features

- InitCapPeriod (e.g., *Mr.*)
- OneCap (e.g., *A*)
- AllCapsPeriod (e.g., *CORP.*)
- ContainDigit (e.g., *AB3, 747*)
- TwoD (e.g., *99*)
- FourD (e.g., *1999*)
- DigitSlash (e.g., *01/01*)
- Dollar (e.g., *US$20*)
- Percent (e.g., *20%*)
- DigitPeriod (e.g., *$US3.20*)

Local Features

- Dictionary word lists
  - Person first names, person last names, organization names, location names
- Person prefix list (e.g., *Mr.*, *Dr.*), corporate suffix list (e.g., *Corp.*, *Inc.*)
  - Obtained from training data

- Month names, Days of the week, Numbers

Global Features

- Initcaps of other occurrences

Even Daily News have made the same mistake ….

They criticised **Daily News** for missing something **even** a boy would have noticed….

Global Features

- Person prefix and corporate suffix of other occurrences

Mary Kay Names Vice Chairman

Richard C. Bartlett was named to the newly created position of vice chairman of **Mary Kay Corp.**

Global Features

- Acronyms

The **Federal Communications Commission** killed

that plan last year … …

The company is still trying to challenge the **FCC**'s earlier decision … …

Global Features

- Sequence of initial caps

  [HL] First Fidelity Unit Heads Named

  [TXT] Both were executive vice presidents at First Fidelity.

NER in Indian Languages

Problems for NER in Indian Languages

- Lacks capitalization information
- More diverse Indian person names
  - Lot of person names appear in the dictionary with other specific meanings
    - For e.g., *KabiTA* (Person name vs. Common noun with meaning 'poem')
- High inflectional nature of Indian languages
  - Richest and most challenging sets of linguistic and statistical features resulting in long and complex wordforms
- Free word order nature of Indian languages
- Resource-constrained environment of Indian languages
  - PoS taggers, morphological analyzers, name lists etc. are not available in the web
- Non-availability of sufficient published works

NER in Indian Languages

- LI and McCallum (2004)-Hindi
  - CRF model using feature induction technique to automatically construct the features
  - Features:
    - Word text, character n-grams (n=2, 3, 4), word prefix and suffix of lengths 2,3,4
    - 24 Hindi gazetteer lists
    - Features at the current, previous and next sequence positions were made available
  - Dataset: 601 BBC and 27 EMI Hindi documents
  - Performance
    - *F-measure* of 71.5% with an early stopping point of 240 iterations of L-BFGS for the 10-fold cross validation experiments

NER in Indian Languages

- Saha et al. (2008)-Hindi
  - ME model
  - Features:
    - Statistical and linguistic feature sets
    - Hindi gazetteer lists
    - Semi-automatic induction of context patterns
    - Context patterns as features of the MaxEnt method
  - Dataset: 243K words of Dainik Jagaran (training)
    25K (test)
  - Performance
    - *F-measure* of 81.52%

NER in Indian Languages

- Patel et al. (2008)-Hindi and Marathi
  - Inductive Logic Programming (ILP) based techniques for automatically extracting rules for NER from tagged corpora and background knowledge
  - Dataset: 54340 (Marathi), 547138 (Hindi)
  - Performance
    - *PER:* 67%, *LOC:* 71% *and ORG:* 53% (Hindi)
    - *PER:* 82%, *LOC:* 48% *and ORG:* 55% (Hindi)
  - Advantages over rule-based system
    - development time reduces by a factor of 120 *compared to a linguist doing the entire rule development*
    - *a complete and consistent view of all significant patterns in the data at the level of abstraction*

NER in Indian Languages

- Ekbal and Saha (2011)-Bengali, Hindi, Telugu and Oriya
  - Genetic algorithm based weighted ensemble
  - Classifiers: ME, CRF and SVM
  - Features:
    - Word text, word prefix and suffix of lengths 1,2,3; PoS
    - Context information, various orthographic features etc.
  - Dataset: Bengali (Training: 312,947; Test: 37,053)

    Hindi (Training: 444,231; Test: 58,682)

    Telugu (Training: 57,179; Test: 4,470)

    Oriya (Training: 93,573; Test: 2,183)
  - Performance
    - *F-measures: Bengali* (92.15%), *Hindi* (92.20%), *Telugu* (84.59%) and *Oriya* (89.26%)

NER in Indian Languages

- Ekbal and Saha (2012)-Bengali, Hindi and Telugu
  - Multiobjective Genetic algorithm based weighted ensemble
  - Classifiers: ME, CRF and SVM
  - Features:
    - Word text, word prefix and suffix of lengths 1,2,3; PoS
    - Context information, various orthographic features etc.
  - Dataset: Bengali (Training: 312,947; Test: 37,053)

    Hindi (Training: 444,231; Test: 58,682)

    Telugu (Training: 57,179; Test: 4,470)

    Oriya (Training: 93,573; Test: 2,183)
  - Performance
    - *F-measures: Bengali* (92.46%), *Hindi* (93.20%), *Telugu* (86.54%)

NER in Indian Languages

- Shishtla et al. (2008)- Telugu and Hindi
  - CRF
  - Character-n gram approach is more effective than word-based model
  - Features
    - Word-internal features, PoS, chunk etc.
    - No external resources

  -Datasets: Telugu (45,714 tokens); Hindi ((45,380 tokens)

  -Performance
    - F-measures: Telugu (49.62%), Hindi (45.07%)

NER in Indian Languages

- Vijayakrishna and Sobha (2008)
  - CRF
  - Tourism domain with 106 hierarchical tags
  - Features
    - Roots of words, PoS, dictionary of NEs, patterns of certain types of NEs (date, time, money etc.) etc
  - Performance
    - 80.44%

NER in Indian Languages

- Saha et al. (2008)- Hindi
  - Maximum Entropy
  - Features
    - Statistical and linguistics features
    - Word clustering
    - Clustering used for feature reduction in Maximum Entropy
- -Datasets: 243K Hindi newspaper “Dainik Jagaran”.

  -Performance
    - F-measures: 79.03% (approximately 7% improvement with Clusters)

Other works in Indian Languages NER

- Gali et al. (2008)-Bengali, Hindi, Telugu and Oriya
  - CRF
- Kumar and Kiran (2008)-Bengali, Hindi, Telugu and Oriya
  - CRF
- Srikanth and Murthy (2008) –Telugu
  - CRF
- Goyal (2008)-Hindi
  - CRF
- Nayan et al. (2008)-Hindi
  - Phonetic matching technique

Other works in Indian Languages NER

- Ekbal et al. (2008)-Bengali
  - CRF
- Saha et al. (2009)-Hindi
  - Semi-supervised approach
- Saha et al. (2010)-Hindi
  - SVM with string based kernel function
- Ekbal and Saha (2010)-Bengali, Hindi and Telugu
  - GA based classifier ensemble selection
- Ekbal and Saha (2011)-Bengali, Hindi and Telugu
  - Multiobjective simulated annealing approach for classifier ensemble

Other works in Indian Languages NER

- Saha et al. (2012)-Hindi and Bengali
  - Comparative techniques for feature reductions

- Ekbal and Saha (2012)-Bengali, Hindi and Telugu
  - Multiobjective approach for feature selection and classifier ensemble

- Ekbal et al. (2012)-Hindi and Bengali
  - Active learning
  - Effective in a resource-constrained environment

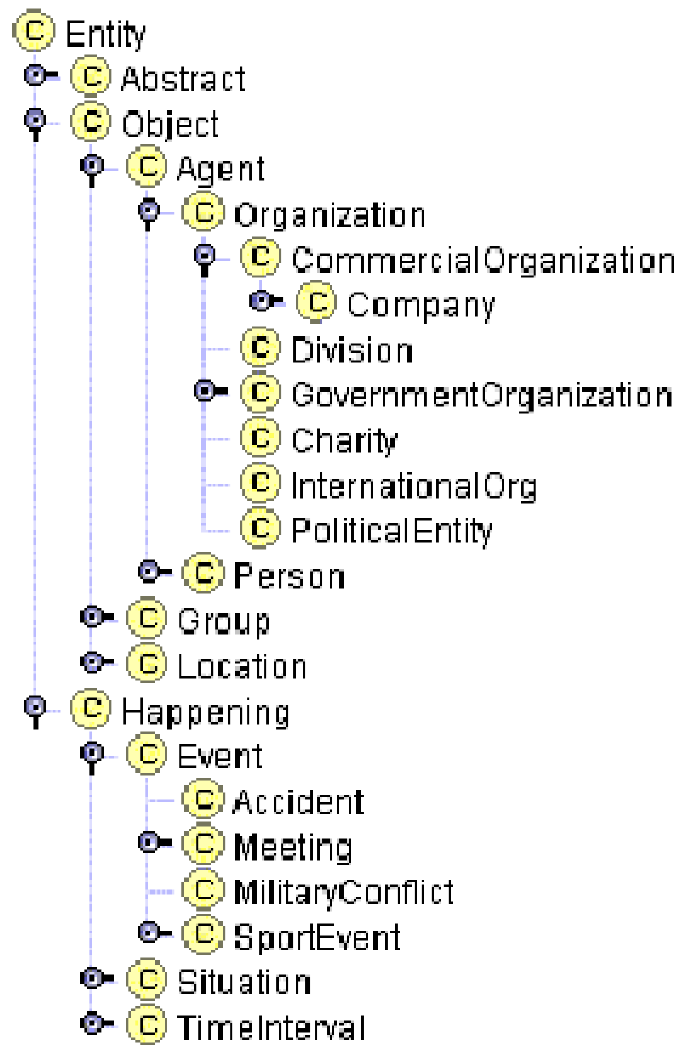
Shared Tasks on Indian Language NER

- NERSSEAL Shared Task- 2008 (http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=2)

- NLPAI ML Contest 2007- (http://ltrc.iiit.ac.in/nlpai_contest07/cgi-bin/index.cgi)

Evaluating Richer NE Tagging

- Hierarchy/ontology-based NE tagging
- Need to take into account distance in the hierarchy
- Tagging a company as a charity is less wrong than tagging it as a person



HMM based NERC

HMM based NERC System (Contd..)

Problem of NE tagging

Let W be a sequence of words

$$W = w_1 , w_2 , \ldots , w_n$$

Let T be the corresponding NE tag sequence

$$T = t_1 , t_2 , \ldots , t_n$$

Task : Find T which maximizes $P (T \mid W)$

$$T' = \text{argmax}_T \, P (T \mid W)$$

HMM based NERC System (Contd..)

By Bayes' Rule,

$$P(T \mid W) = P(W \mid T) * P(T) / P(W)$$

$$T' = \text{argmax}_T \, P(W \mid T) * P(T)$$

- **Models**
  - Fisrt order model (Bigram): The probability of a tag depends only on the previous tag
  - Second order model (Trigram): The probability of a tag depends on the previous two tags

- **Transition Probability**

Bigram → $P(T) = P(t_1) * P(t_2 \mid t_1) * P(t_3 \mid t_2) \ldots\ldots * P(t_n \mid t_{n-1})$

Trigram → $P(T) = P(t_1) * P(t_2 \mid t_1) * P(t_3 \mid t_1\, t_2) \ldots\ldots * P(t_n \mid t_{n-2}\, t_{n-1})$

$$P(T) = P(t_1 \mid \$) * P(t_2 \mid \$\, t_1) * P(t_3 \mid t_1\, t_2) \ldots\ldots * P(t_n \mid t_{n-2}\, t_{n-1})$$

Where, \$ → dummy tag used to represent the beginning of a sentence

HMM based NERC System (Contd..)

- Estimation of unigram, bigram and trigram probabilities from the training corpus

Unigram	:	$P(t_3) = \frac{freq(t_3)}{N}$
Bigram	:	$P(t_3 \mid t_2) = \frac{freq(t_2, t_3)}{freq(t_2)}$
Trigram	:	$P(t_3 \mid t_1, t_2) = \frac{freq(t_1, t_2, t_3)}{freq(t_1, t_2)}$

- Emission Probability

$$P(W \mid T) \approx P(w_1 \mid t_1) * P(w_2 \mid t_2) * \ldots * P(w_n \mid t_n)$$

Emission Probability: $P(w_i \mid t_i) = \frac{freq(w_i, t_i)}{freq(t_i)}$

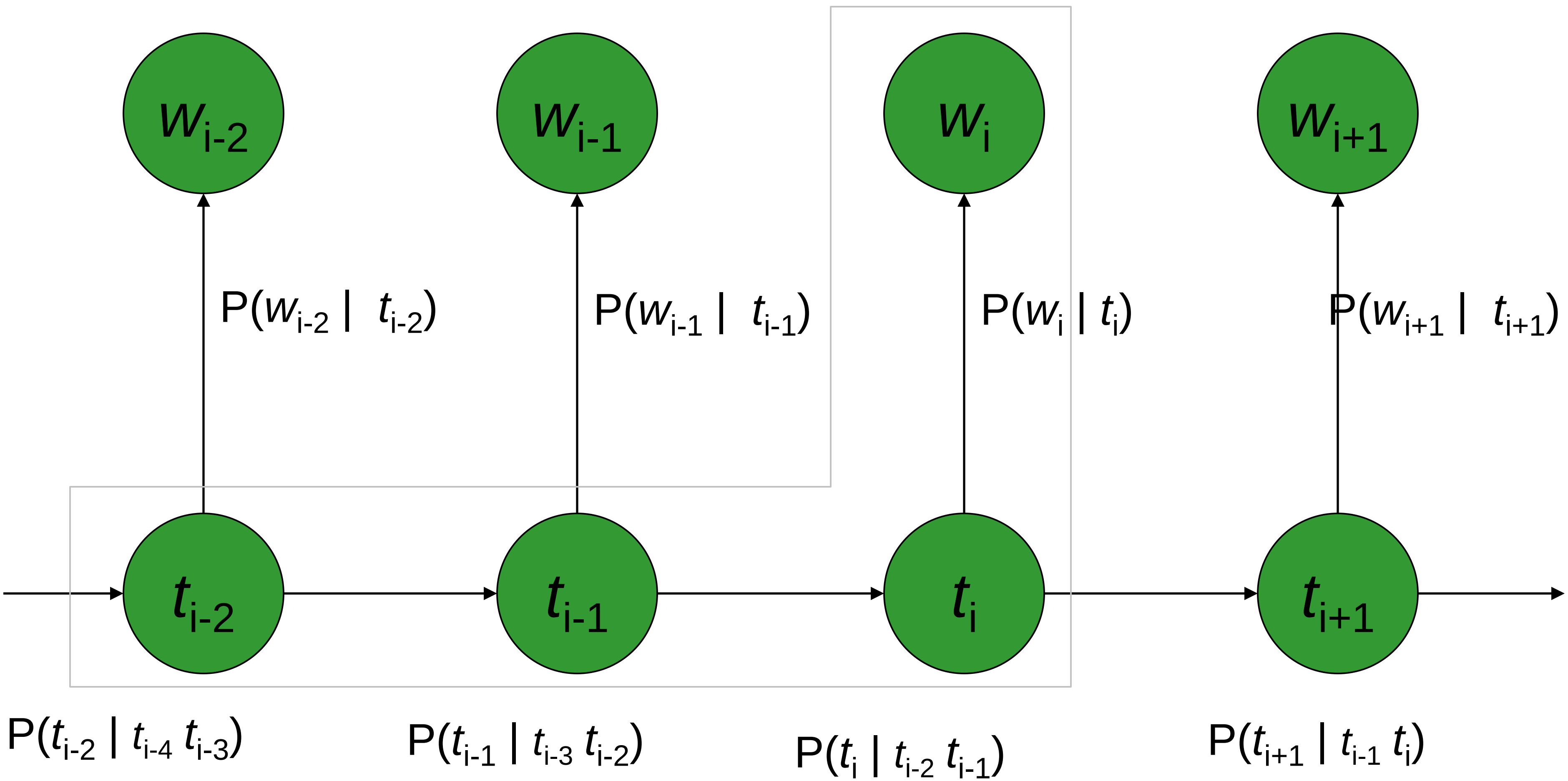
HMM based NERC System (Contd..)

- Context Dependency (Our Modification)
  - Markov model is made more powerful by introducing 1st order context dependent feature

$$P(W \mid T) \approx P(w_1 \mid \$, t_1) * P(w_2 \mid t_1, t_2) * \ldots * P(w_n \mid t_{n-1}, t_n)$$

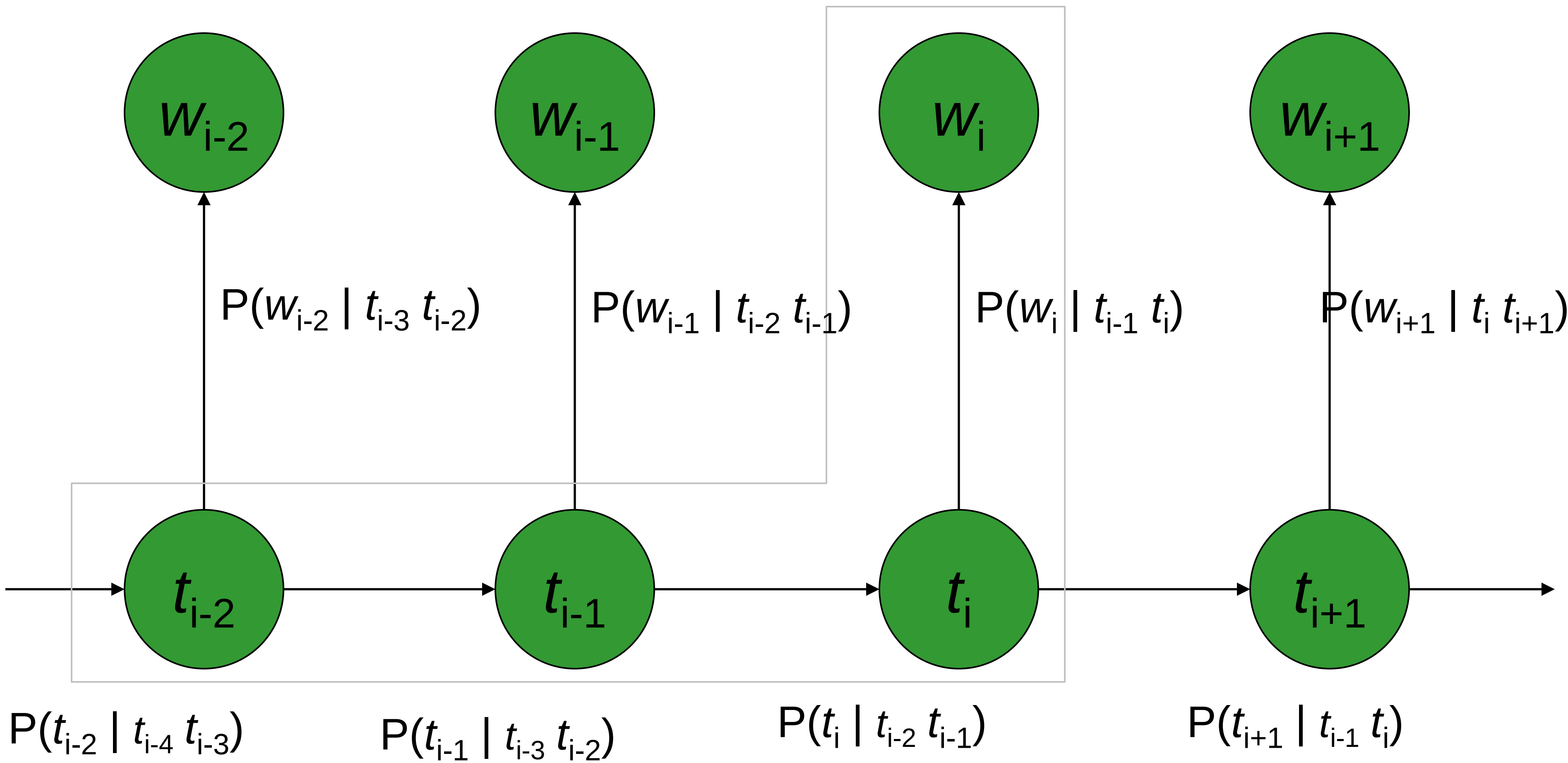
$$P(w_i \mid t_{i-1}, t_i) = \frac{freq(t_{i-1}, t_i, w_i)}{freq(t_{i-1}, t_i)}$$

HMM based NERC System (Contd..)



2nd order Hidden Markov Model

HMM based NERC System (Contd..)



2nd order Hidden Markov Model (Proposed)

HMM based NERC System (Contd..)

- Why Smoothing?
  - Limited training corpus
  - Insufficient instances for each *bigram* or *trigram* to reliably estimate the probability
  - Setting a probability to zero has an undesired effect
- Procedure (*Linear Interpolation*)
  - Transition probability

$$P'(t_n | t_{n-2}, t_{n-1}) = \lambda_1 P(t_n) + \lambda_2 P(t_n | t_{n-1}) + \lambda_3 P(t_n | t_{n-2}, t_{n-1})$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

  - Emission probability

$$P'(w_i | t_{i-1}, t_i) = \theta_1 P(w_i | t_i) + \theta_2 P(w_i | t_{i-1}, t_i)$$

$$\theta_1 + \theta_2 = 1$$

  - Calculation of λ s and Θs (Brants, 2000)

HMM based NERC System (Contd..)

- Handling of unknown words
  - Viterbi algorithm (Viterbi, 1967) attempts to assign a tag to the unknown words
  - $P(w_i \mid t_i) \rightarrow P(f_i \mid t_i)$
    - Calculated based on the features of unknown word
    - Suffixes: Probability distribution of a particular suffix with respect to specific NE tags is generated from all words in the training set that share the same suffix
      - →Variable length person name suffixes (e.g., - *bAbu*[-babu], -*dA* [-da] , -*di*[-di] etc)
      - →Variable length location name suffixes (e.g., -*lYAnd*[-land], -*pur*[pur], -*liYA*[-lia]) etc)

Results of the HMM based System: Bengali

Model	Reacall (in %)	Precision (in %)	F-Score (in %)
HMM (*bigram*)	76.92	74.79	75.84
HMM (*trigram*)	77.33	75.98	76.65

Results on development set

Observation:

1. Second order model performs better than first order model with a margin of 0.81%

2. Trigram selected to report the test set results

Model	Reacall (in %)	Precision (in %)	F-Score (in %)
Baseline (i.e., Model A)	64.32	67.29	65.77
HMM	77.04	75.17	75.76

Results on the test set

Observation: HMM performs better than the *baseline* model with more than 12.72%, 7.88%, and 9.99% in *Recall*, *Precision*, and *F-Score* values, respectively

Ensemble Learning: A brief Introduction

Drawbacks of Single Classifier

- The "best" classifier not necessarily the ideal choice

- For solving a classification problem, many individual classifiers with different parameters are trained
  - The "best" classifier will be selected according to some criteria e.g., *training accuracy* or *complexity of the classifiers*

- Problems: Which one is the best?
  - Maybe more than one classifiers meet the criteria (e.g. same training accuracy), especially in the following situations:
    - -Without sufficient training data
    - – Learning algorithm leads to different local optima easily

Drawbacks of Single Classifier

- Potentially valuable information may be lost by discarding the results of less-successful classifiers

  E.g., the discarded classifiers may correctly classify some samples

- Other drawbacks
  - Final decision must be wrong if the output of selected classifier is wrong
  - Trained classifier may not be complex enough to handle the problem

Ensemble Learning

- Employ multiple learners and combine their predictions

- Methods of combination:
  - Bagging, boosting, voting
  - Error-correcting output codes
  - Stacked generalization
  - Cascading
  - …
- **Advantage**: improvement in predictive accuracy

- **Disadvantage**: it is difficult to understand an ensemble of classifiers

Evolutionary Algorithms for Classifier Ensemble

Genetic Algorithm: **Quick Overview**

- Randomized search and optimization technique
- Evolution produces good individuals, similar principles might work for solving complex problems
- Developed: USA in the 1970' s by J. Holland
- Got popular in the late 1980's
- Early names: J. Holland, K. DeJong, D. Goldberg
- Based on ideas from *Darwinian Evolution*
- Can be used to solve a variety of problems that are not easy to solve using other techniques

Genetic Algorithm: Similarity with Nature

Genetic Algorithms	←→	Nature
A solution (phenotype)		Individual
Representation of a solution (*genotype*)		Chromosome
Components of the solution		Genes
Set of solutions		Population
Survival of the fittest (*Selection*)		Darwins theory
Search operators		Crossover and mutation
Iterative procedure		Generations

Basic Steps of Genetic Algorithm

1. $t = 0$
2. initialize population $P(t)$ /* $Popsize = |P|$ */
3. for $i = 1$ to $Popsize$
   compute fitness $P(t)$
4. $t = t + 1$
5. if termination criterion achieved go to step 10
6. select (P)
7. crossover (P)
8. mutate (P)
9. go to step 3
10. output best chromosome and stop

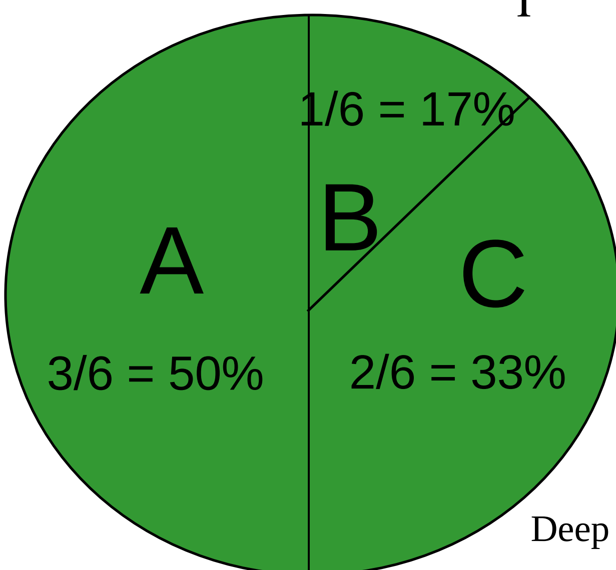
End

Example population

No.	Chromosome	Fitness
1	1010011010	1
2	1111100001	2
3	1011001100	3
4	1010000000	1
5	0000010000	3
6	1001011111	5
7	0101010101	1
8	1011100111	2

GA operators: **Selection**

- Main idea: better individuals get higher chance
  - Chances proportional to fitness
  - Implementation: roulette wheel technique
    - » Assign to each individual a part of the roulette wheel
    - » Spin the wheel n times to select n individuals



fitness(A) = 3

fitness(B) = 1

fitness(C) = 2

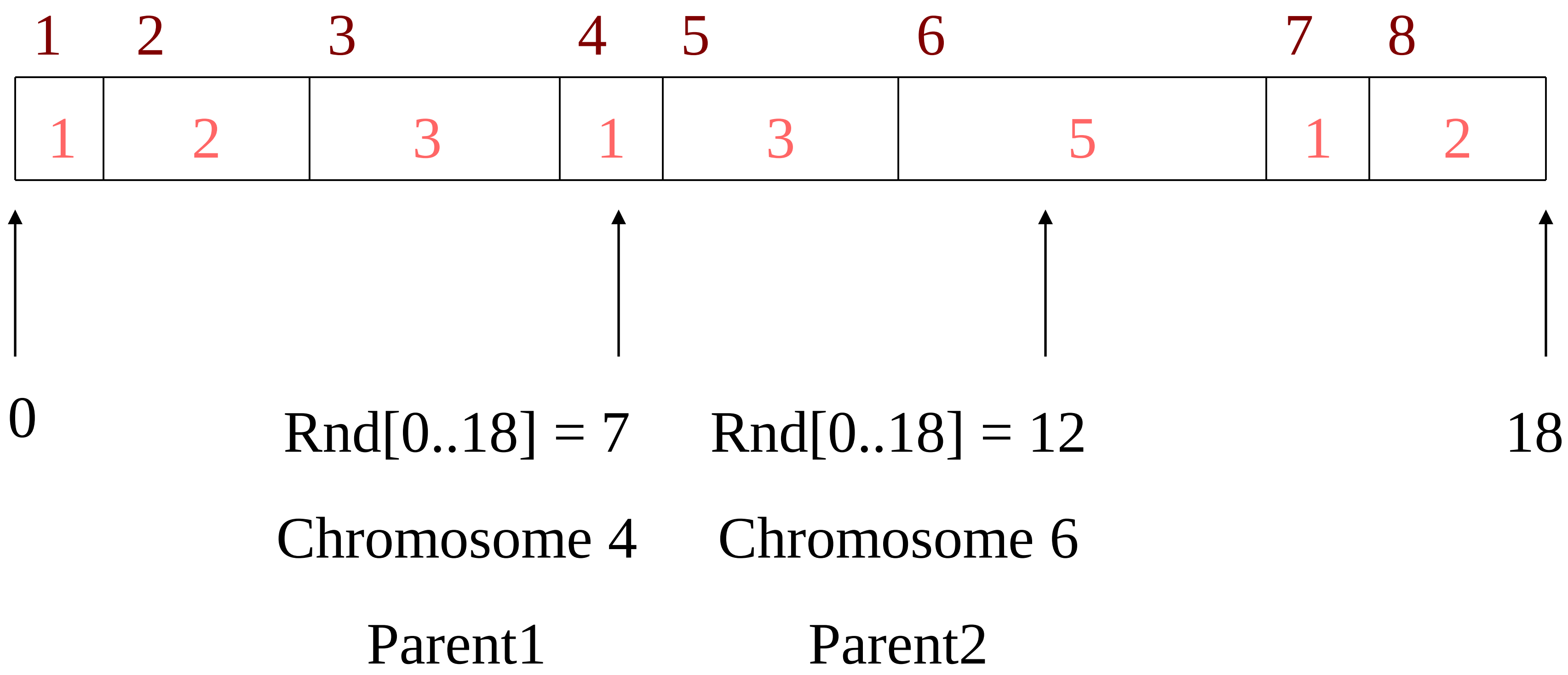
GA operator: **Selection**

– Add up the fitness's of all chromosomes

– Generate a random number R in that range

– Select the first chromosome in the population that - when all previous fitness's are added including the current one- gives you at least the value R

Roulette Wheel Selection



GA operator: **Crossover**

- Choose a random point on the two parents
- Split parents at this crossover point
- With some high probability (*crossover rate*) apply crossover to the parents
  - P_c typically in range (0.6, 0.9)
- Create children by exchanging tails

Crossover: Recombination

1010000000 Parent1

1001011111 Parent2

Offspring1 1011011111

Offspring2 1000000000

Crossover single point - random

Single Point Crossover

n-point crossover

- Choose n random crossover points
- Split along those points
- Glue parts, alternating between parents
- Generalisation of 1 point (still some positional bias)

Mutation

mutate

	Original offspring	Mutated offspring
Offspring1	1011011111	1011001111
Offspring2	1010000000	1000000000

With some small probability (the *mutation rate*) flip each bit in the offspring (*typical values between 0.1 and 0.001*)

A. Ekbal and S. Saha (2011). Weighted Vote-Based Classifier Ensemble for Named Entity Recognition: A Genetic Algorithm-Based Approach. ACM Transactions on Asian Language Information Processing (ACM TALIP), Vol. 2(9),

DOI=10.1145/1967293.1967296

http://doi.acm.org/10.1145/1967293.1967296

Weighted Vote based Classifier Ensemble

- Motivation
  - All classifiers are not equally good at detecting all the classes

- Weighted voting: weights of voting vary among the classes for each classifier
  - *High*: Classes for which the classifier perform good
  - *Low*: Classes for which it's output is not very reliable
- *Crucial issue*: Selection of appropriate weights of votes per classifier

Problem Formulation

Let *no. of classifiers*=N, and *no. of classes*=M

Find the weights of votes V per classifier optimizing a function F(V)

-V: an real array of size N × M

-V(i , j) : weight of vote of the *i*th classifier for the *j*th class

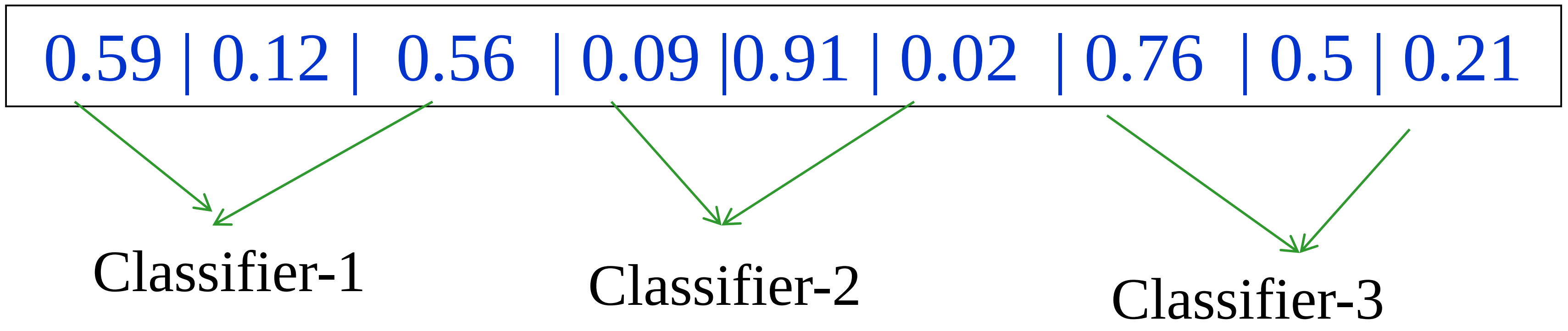
-V(i , j) ε [0, 1] denotes the degree of confidence of the *i*th classifier for the *j*th class

maximize F(B) ;

F ε {recall, precision, F-measure} and B is a subset of A

Here, *F1= F-measure*

Chromosome representation



- Real encoding used
- Entries of chromosome randomly initialized to a real (r) between 0 and 1: r = rand () / RAND_MAX+1
- If the population size P then all the P number of chromosomes of this population are initialized in the above way

Fitness Computation

Step-1: For M classifiers, F_i $i = 1$ to M be the F-measure values

Step-2: Train each classifier with 2/3 training data and evaluate with the 1/3 part

Step-3: For ensemble output of the 1/3 test data, apply weighted voti outputs of M classifiers

(a). Weight of the output label provided by the *mth* classifier = I (m, i)

Here, *I(m, i) is the entry of the chromosome corresponding to mth* classifier a

(b). Combined score of a class for a word *w*

$$f(c_i) = \sum I(m,i) \times F_m, \quad \forall m = 1 \text{ to } M \text{ and } op(w,m) = c_i$$

Fitness Computation

Op(w, m): output class produced by the *mth* classifier for word *w*

Class receiving the maximum score selected as joint decision

Step-4: Compute overall F-measure value for 1/3 data

Step-5: Steps 3 and 4 repeated to perform 3-fold cross validation

Step-6: Objective function or fitness function = F-measure_{avg}

Objective: Maximize the objective function using search capability of GA

Other Parameters

- Selection
  – Roulette wheel selection (*Holland, 1975; Goldberg, 1989*)
- Crossover
  – Normal Single-point crossover (Holland, 1975)
- Mutation
  – Probability selected adaptively (*Srinivas and Patnaik, 1994*)
  – Helps GA to come out from local optimum

Termination Condition

- Execute the processes of *fitness computation*, *selection*, *crossover*, and *mutation* for a maximum number of generations
- *Best solution*-Best string seen up to the last generation

- Best solution indicates
  - Optimal voting weights for all classes in each classifier

- Elitism implemented at each generation
  - Preserve the best string seen up to that generation in a location outside the population
  - Contains the most suitable classifier ensemble

NE Features: Mostly language independent

NE Features

- Context Word: Preceding and succeeding words
- Word Suffix
  - Not necessarily linguistic suffixes
  - Fixed length character strings stripped from the endings of words
  - Variable length suffix -binary valued feature
- Word Prefix
  - Fixed length character strings stripped from the beginning of the words
- Named Entity Information: Dynamic NE tag (s) of the previous word (s)

NE Features

- First Word (binary valued feature): Check whether the current token is the first word in the sentence

- Length (binary valued): Check whether the length of the current word less than three or not (shorter words rarely NEs)

- Position (binary valued): Position of the word in the sentence

- Infrequent (binary valued): Infrequent words in the training corpus most probably NEs

NE Features

- Digit features: Binary-valued
  - Presence and/or the exact number of digits in a token
    - CntDgt : Token contains digits
    - FourDgt: Token consists of four digits
    - TwoDgt: Token consists of two digits
    - CnsDgt: Token consists of digits only

- Combination of digits and punctuation symbols
  - CntDgtCma: Token consists of digits and comma
  - CntDgtPrd: Token consists of digits and periods

NE Features

- Combination of digits and symbols
  - CntDgtSlsh: Token consists of digit and slash
  - CntDgtHph: Token consists of digits and hyphen
  - CntDgtPrctg: Token consists of digits and percentages

- Combination of digit and special symbols
  - CntDgtSpl: Token consists of digit and special symbol such as $, # etc.

NE Features

- Part of Speech (POS) Information: POS tag(s) of the current and/or the surrounding word(s)
  - SVM-based POS tagger (Ekbal and Bandyopadhyay, 2008)
  - SVM based NERC→POS tagger developed with a fine-grained tagset of 27 tags
  - Coarse-grained POS tagger
    - Nominal, PREP (Postpositions) and Other

- Gazetteer based features (binary valued): Several features extracted from the gazetteers

Datasets

- Web-based Bengali news Corpus (Ekbal and Bandyopadhyay, 2008, *Language Resources and Evaluation of Springer*)
  - *34 million* wordforms
  - News data collection of 5 years

- NE annotated corpus for Bengali
  - Manually annotated 250K wordforms
  - IJCNLP-08 Shared Task on NER for South and South East Asian Languages (available at http://ltrc.iiit.ac.in/ner-ssea-08)
- NE annotated datasets for Hindi and Telugu
  - NERSSEAL shared task

NE Tagset

- Reference Point- CoNLL 2003 shared task tagset
- Tagset: 4 NE tags
  - Person name
  - Location name
  - Organization name
  - Miscellaneous name (*date*, *time*, *number*, *percentages*, *monetary expressions* and *measurement expressions*)

- IJCNLP-08 NERSSEAL Shared Task Tagset: Fine-grained 12 NE tags (available at http://ltrc.iiit.ac.in/ner-ssea-08)

- Tagset Mapping (12 NE tags→4 NE tags)
  - ❑ NEP → Person name
  - ❑ NEL→ Location name
  - ❑ NEO→ Organization name
  - ❑ NEN [number], NEM [Measurement] and NETI [time]→Miscellaneous name
  - ❑ NETO [title-object], NETE [term expression], NED [designations], NEA [abbreviations], NEB [brand names], NETP [title persons

Training and Test Datasets

Language	#Words in training	#NEs in training	#Words in test	#NEs in test
Bengali	312,947	37,009	37,053	4,413
Hindi	444,231	26,432	32,796	58,682
Telugu	57,179	4,470	6,847	662
Oriya	93,573	4,477	2,183	206

Experiments

- Classifiers used
  - Maximum Entropy (ME): Java based OpenNLP package (http://maxent.sourceforge.net/)
  - Conditional Random Field: C++ based CRF++ package (http://crfpp.sourceforge.net/)
  - Support Vector Machine:
    - YamCha toolkit

      (http://chasen-org/ taku/software/yamcha/)
    - TinySVM-0.07

      (http://cl.aist-nara.ac.jp/ taku-ku/software/TinySVM)
    - Polynomial kernel function

Experiments

- GA: population size=50, number of generations=40, mutation and crossover probabilities are selected adaptively.

- Baselines
  - Baseline 1: Majority voting of all classifiers
  - Baseline 2: Weighted voting of all classifiers (*weight*: overall average F-measure value)
  - Baseline 3: Weighted voting of all classifiers (*weight*: F-measure value of the individual class)

Results (*Bengali*)

Model	Recall	Precision	F-measure
Best Individual Classifier	89.42	90.55	89.98
Baseline-1	84.83	85.90	85.36
Baseline-2	85.25	86.97	86.97
Baseline-3	86.97	87.34	87.15
Stacking	90.17	91.74	90.95
ECOC	89.78	90.89	90.33
QBC	90.01	91.09	90.55
GA based ensemble	92.08	92.22	92.15

Results (*Hindi*)

Model	Recall	Precision	F-measure
Best Individual Classifier	88.72	90.10	89.40
Baseline-1	63.32	90.99	74.69
Baseline-2	74.67	94.73	83.64
Baseline-3	75.52	96.13	84.59
Stacking	89.80	90.61	90.20
ECOC	90.16	91.11	90.63
GA based ensemble	96.07	88.63	92.20

Results (*Telugu*)

Model	Recall	Precision	F-measure
Best Individual Classifier	77.42	77.99	77.70
Baseline-1	60.12	87.39	71.23
Baseline-2	71.87	92.33	80.33
Baseline-3	72.22	93.10	81.34
Stacking	77.65	84.12	80.76
ECOC	77.96	85.12	81.38
GA based ensemble	78.82	91.26	84.59

Results (*Oriya*)

Model	Recall	Precision	F-measure
Best Individual Classifier	86.55	88.03	87.29
Baseline-1	86.95	88.33	87.63
Baseline-2	87.12	88.50	87.80
Baseline-3	87.62	89.12	88.36
Stacking	87.90	89.53	88.71
ECOC	87.04	88.56	87.79
GA based ensemble	88.56	89.98	89.26

Results (*English*)

Model	Recall	Precision	F-measure
Best Individual Classifier	86.16	85.24	86.31
Baseline-1	85.75	86.12	85.93
Baseline-2	86.20	87.02	86.61
Baseline-3	86.65	87.25	86.95
Stacking	85.93	86.45	86.18
ECOC	86.12	85.34	85.72
GA based ensemble	88.72	88.64	88.68

Current Trends in NE Research

- Development of domain-independent and language-independent systems
  - Can be easily portable to different domains and languages

- Fine-grained NE classification
  - May be at the hierarchy of WordNet
  - Beneficial to the fine-grained IE
  - Helps in Ontology learning

Current Trends in NE Research

- NER systems in non-newswire domains
  - Humanities (arts, history, archeology, literature etc.): *lots of non-traditional entities are present*
  - Chemical and bio-chemical (*long and nested NEs*)
  - Biomedical texts and clinical records (*long and nested NEs; does not follow any standard nomenclature)*
  - Unstructured datasets such as Twitter, online product reviews, blogs, SMS etc.

Study Materials: References

- ***Named Entities: Recognition, Classification and Use, Special Issue of Lingvisticae Investigationes Journal***, Satoshi Sekine and Elisabete Ranchhod (Eds.), Vol. 30:1 (2007), John Benjamins Publishing Company
- All relevant conferences- ACL, COLING, EACL, IJCNLP, CiCLing , AAAI, ECAI etc.
- Named Entities Workshop (NEWS)
- Biotext Mining challenges- BioCreative, BioNLP etc.
- NER in unstructured text: NER in twitter (*ACL 2015 and COLING 2016 Shared Tasks*), NER in code-mixed data (*Fire shared task-16*)

Important Resources

- Stanford NER: Classifier: CRF; Language: English; Types: PER, LOC and ORG
- LingPipe: Hybrid; News Entities: PER, LOC and ORG; Biomedical: Genes, Organisms, Chemicals
- TextPro: Supervised SVM (YamCha); Languages: Italian, English and German; Entities: PER, LOC and ORG
- GATE: Hybrid System; Language: English; Entities: PER, LOC and ORG
- BANNER: Classifier: CRF; Entities: Gene and Gene Products
- GENIA Tagger: HMM; Entities: Protein, DNA, RNA, Cell_Line and Cell_Type
- Important Datasets: CoNLL 2002/2003, JNLPBA-2004, BioCreative, IJCNLP-08 NERSSEAL, Twitter NER (W-NUT 2016/15)

Thank you for your attention