Named Entity Recognition and Classification

Asif Ekbal

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

AI-NLP-ML @IIT Patna

- Faculty
  - Prof. Pushpak Bhattacharyya
  - Dr. Asif Ekbal
  - Dr. Sriparna Saha
- PhD Students/Research Engineers/Linguists
  - 16
- M.Tech/B.Tech Students
  - 18

Thrust Areas: AI, NLP and ML

- Machine Translation (E-IL, IL-E)
- Cross-lingual Search
- Sentiment Analysis (Product reviews, Cross-lingual and Multilingual, Code-mixed environment)
- Information Extraction-NER, Coreference Resolution, Relation Extraction etc.

- AI and Machine Learning for Health Care
- Text Mining in Biomedicine
- Bioinformatics

Research Labs on AI, NLP & ML- Industry Sponsored

- Centre for Excellence of Natural Language Processing-Reed Elsevier Publishing Company
  - Some novel R&D problems in the broad areas of AI, NLP and ML
- ATL IIT AI Lab- Accenture Pvt. Limited
  - Research on QA, Multilingual Support and Virtual Agent
- Shusrut: ezDI Research Lab on Health Informatics-ezDI (Ahmadabad, India; Head office-USA)
- Research on MT, Sentiment Analysis–Process Nine Technologies

Visiting Professors: Through GIAN

- Prof. Sadao Kurohasi- Kyoto University, Japan
  - Offered a course on NLP in summer (May 2-8, 2016)

- Prof. Carlos A Coello Coello- CINVESTAV/IPN, Mexico

  -Offered a course on Multi-objective Optimization during December 15-22, 2016

Outline

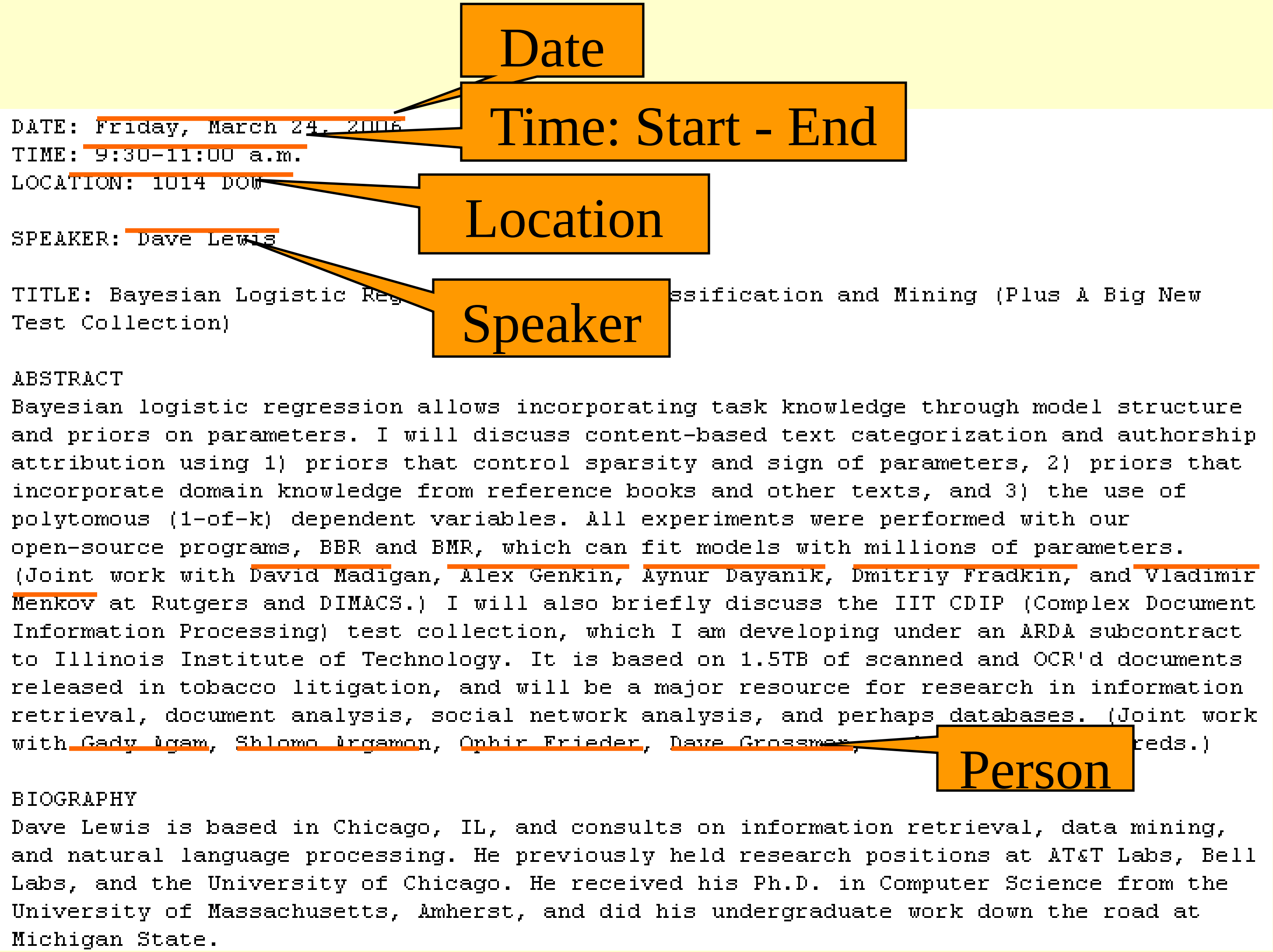
- NERC
  - Background
  - Introduction to the various issues of NERC
  - NERC in different Languages
  - NERC in Indian Languages
- Bio-Text Mining
  - Introduction
  - NE Extraction in Biomedicine

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

SPEAKER: Dave Lewis

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.)

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.

Date

Time: Start - End

Location

Speaker

Person

What is “Information Extraction”

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

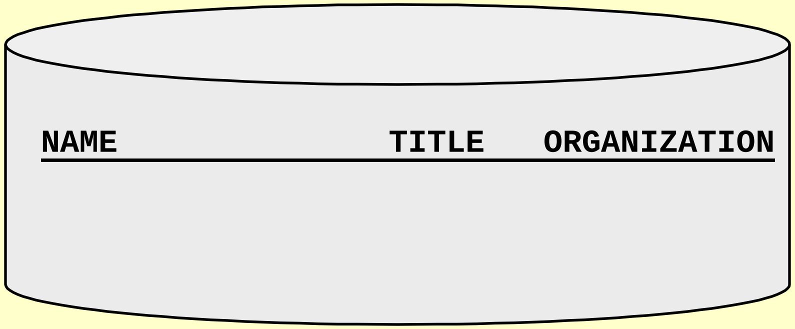
October 14, 2002, 4:00 a.m. PT

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 tex**

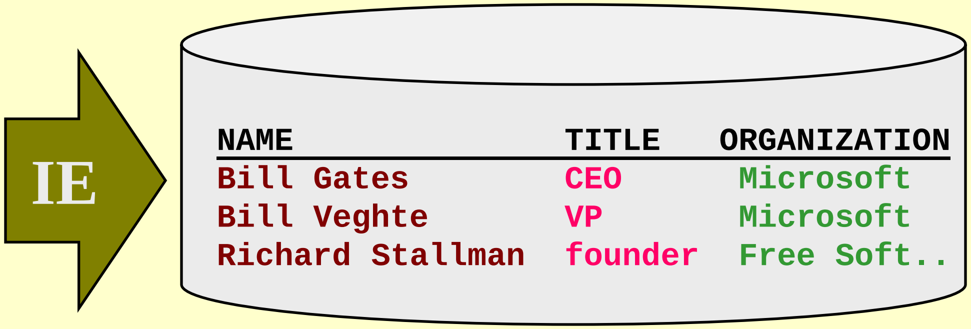
October 14, 2002, 4:00 a.m. PT

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…



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

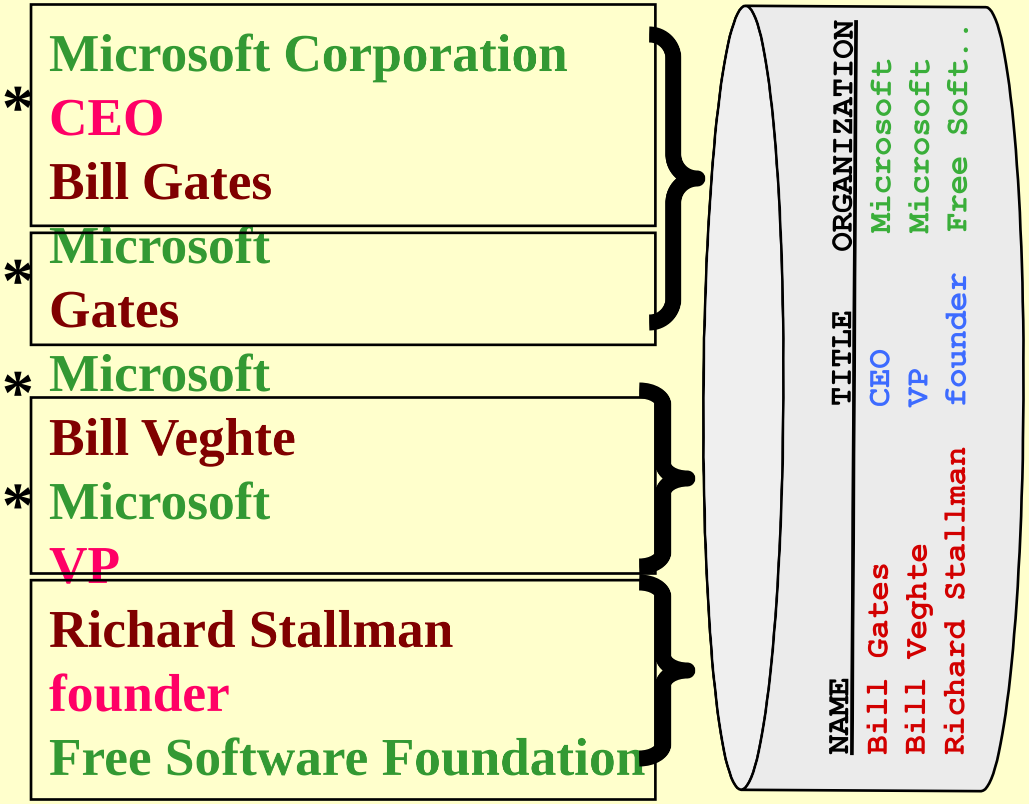
October 14, 2002, 4:00 a.m. PT

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…



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** 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 and Biochemical corpora
  - BioNLP-04 shared task
  - BioCreative shared tasks
  - AiMed
- NER in Tweet
  - ACL-IJCNLP Workshop on Noisy User-generated Text (W-NUT)

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 = $(\beta^2 + 1)PR \,/\, \beta^2R + P$

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

The Evaluation Metric (2)

Precision =

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

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

Pre-processing for NER

- Format detection
- Word segmentation (for languages like Chinese)
- Tokenisation
- Sentence splitting
- Part-of-Speech (PoS) tagging

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

List lookup approach-baseline

- System that recognises only entities stored in its lists (gazetteers)

- Advantages - Simple, fast, language independent, easy to retarget (just create lists)

- Disadvantages - collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Shallow Parsing Approach (internal structure)

- Internal evidence–names often have internal structure. These components can be either stored or guessed,

e.g. location:

- Cap. Word + {City, Forest, Centre, River}

  e.g. Sundarban Forest

- Cap. Word +{Street, Boulevard, Avenue, Crescent, Road}

  e.g. MG Road

e.g. Person

- Word + {Kumar, Chandra} + Word

  E.g., Naresh Kumar Singh

Problems with the shallow parsing approach

- Ambiguously capitalized words (first word in sentence)
  [All American Bank] vs. All [State Police]

- Semantic ambiguity
  "John F. Kennedy" = airport (location)
  "Philip Morris" = organization

- Structural ambiguity
  [Cable and Wireless] vs. [Microsoft] and [Dell];
  [Center for Computational Linguistics] vs. message from
  [City Hospital] for [John Smith]

Shallow Parsing Approach with Context

- Use of context-based patterns is helpful in ambiguous cases

- "David Walton" and "Goldman Sachs" are indistinguishable

- But with the phrase "David Walton of Goldman Sachs" and the Person entity "David Walton" recognised, we can use the pattern "[Person] of [Organization]" to identify "Goldman Sachs“ correctly

Examples of context patterns

- [PERSON] earns [MONEY]
- [PERSON] joined [ORGANIZATION]
- [PERSON] left [ORGANIZATION]
- [PERSON] joined [ORGANIZATION] as [JOBTITLE]
- [ORGANIZATION]'s [JOBTITLE] [PERSON]
- [ORGANIZATION] [JOBTITLE] [PERSON]
- the [ORGANIZATION] [JOBTITLE]
- part of the [ORGANIZATION]
- [ORGANIZATION] headquarters in [LOCATION]
- price of [ORGANIZATION]
- sale of [ORGANIZATION]
- investors in [ORGANIZATION]
- [ORGANIZATION] is worth [MONEY]
- [JOBTITLE] [PERSON]
- [PERSON], [JOBTITLE]

Caveats

- Patterns are only indicators based on likelihood
- Can set priorities based on frequency thresholds
- Need training data for each domain
- More semantic information would be useful (e.g. to cluster groups of verbs)

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

Gazetteer lists for rule-based NER

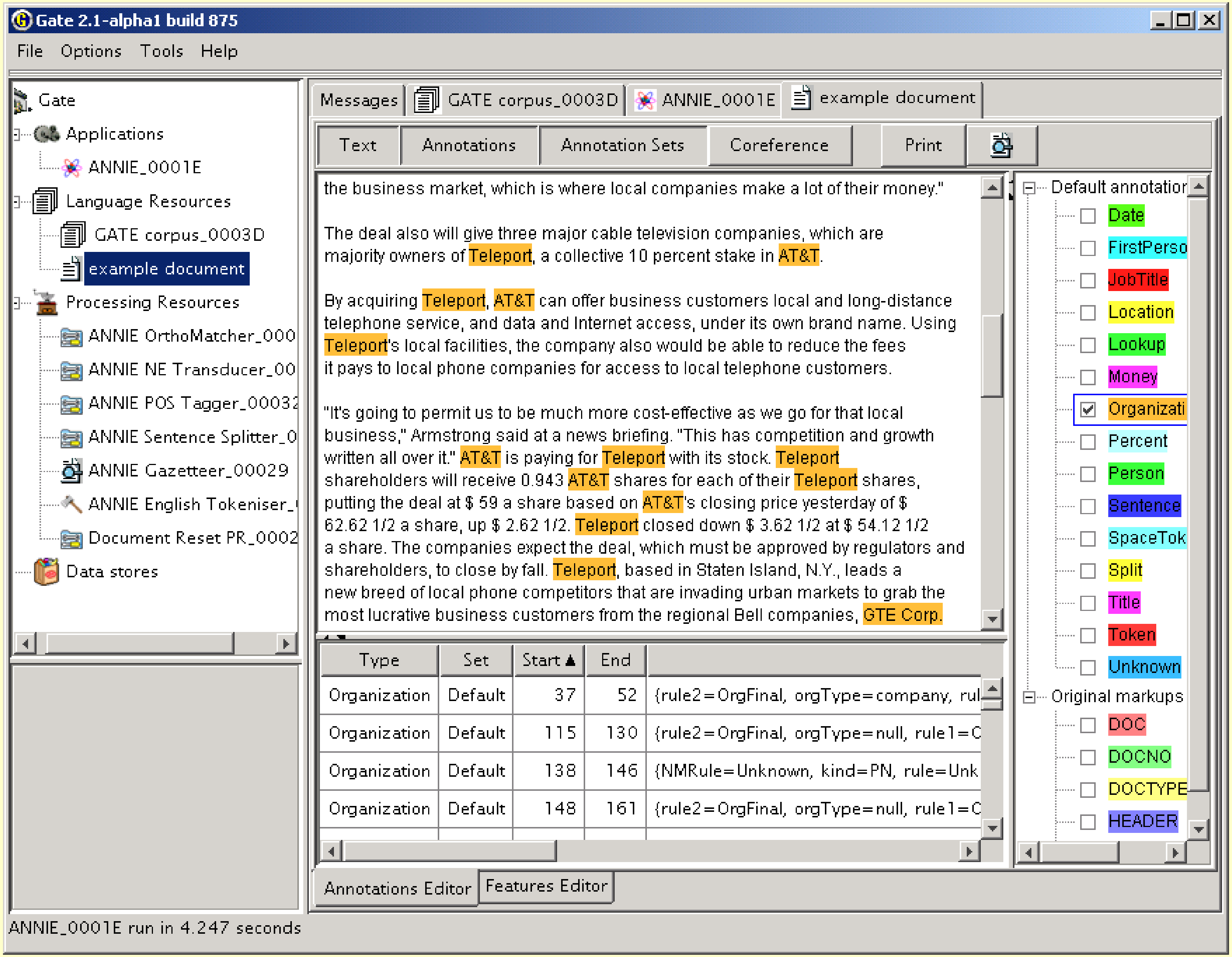
- Needed to store the indicator strings for the internal structure and context rules

- Internal location indicators – e.g., {*river, mountain, forest*} for natural locations; {*street, road, crescent, place, square*, …} for address locations

- Internal organisation indicators–e.g., company designators {*GmbH, Ltd, Inc*, …}

- Produces Lookup results of the given kind

Named Entities in GATE



the business market, which is where local companies make a lot of their money."

The deal also will give three major cable television companies, which are majority owners of Teleport, a collective 10 percent stake in AT&T.

By acquiring Teleport, AT&T can offer business customers local and long-distance telephone service, and data and Internet access, under its own brand name. Using Teleport's local facilities, the company also would be able to reduce the fees it pays to local phone companies for access to local telephone customers.

"It's going to permit us to be much more cost-effective as we go for that local business," Armstrong said at a news briefing. "This has competition and growth written all over it." AT&T is paying for Teleport with its stock. Teleport shareholders will receive 0.943 AT&T shares for each of their Teleport shares, putting the deal at $ 59 a share based on AT&T's closing price yesterday of $ 62.62 1/2 a share, up $ 2.62 1/2. Teleport closed down $ 3.62 1/2 at $ 54.12 1/2 a share. The companies expect the deal, which must be approved by regulators and shareholders, to close by fall. Teleport, based in Staten Island, N.Y., leads a new breed of local phone competitors that are invading urban markets to grab the most lucrative business customers from the regional Bell companies, GTE Corp.

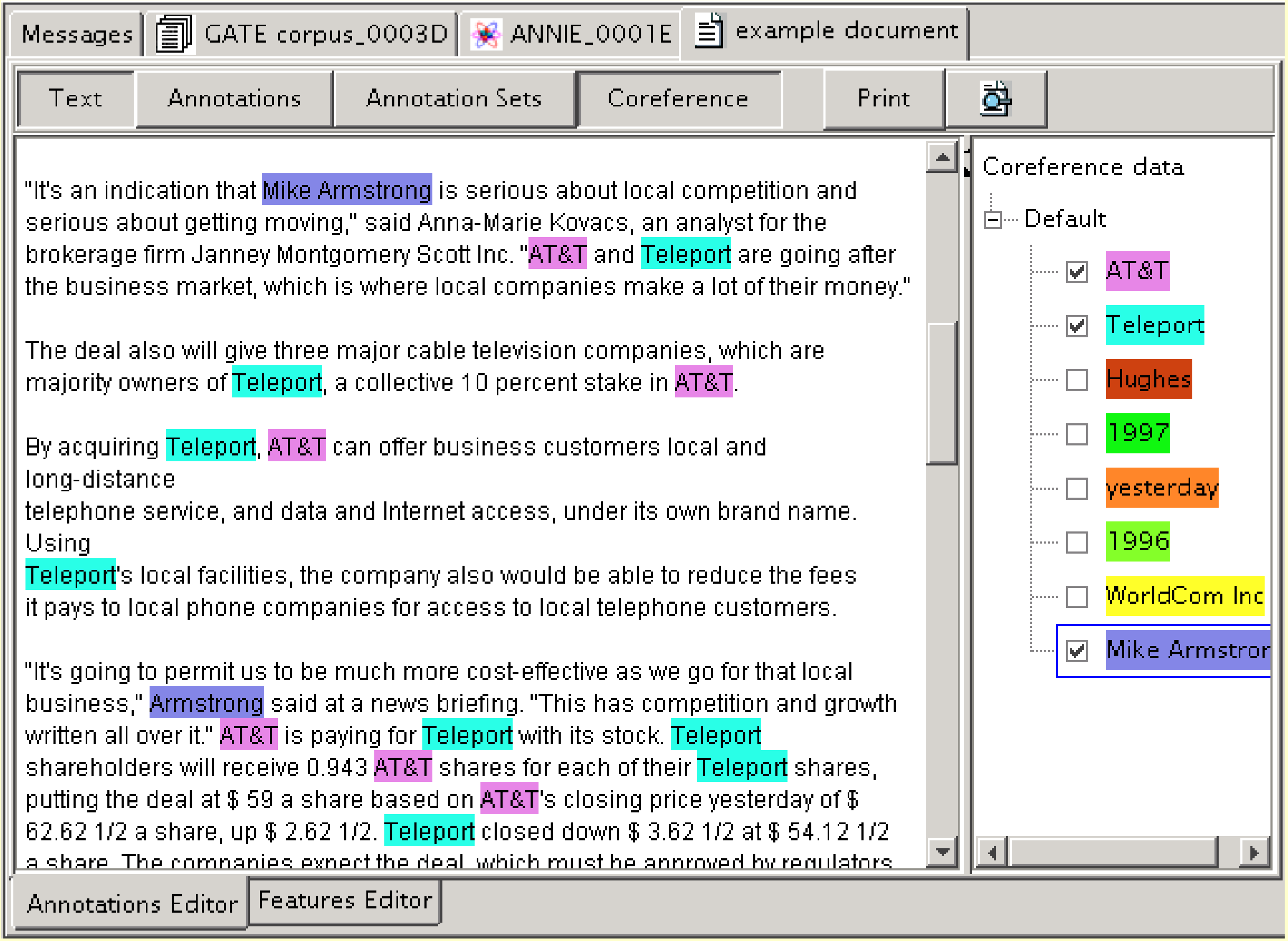
Type	Set	Start ▲	End	
Organization	Default	37	52	{rule2=OrgFinal, orgType=company, rul
Organization	Default	115	130	{rule2=OrgFinal, orgType=null, rule1=C
Organization	Default	138	146	{NMRule=Unknown, kind=PN, rule=Unk
Organization	Default	148	161	{rule2=OrgFinal, orgType=null, rule1=C

ANNIE_0001E run in 4.247 seconds

Using co-reference to classify ambiguous NEs

- Orthographic co-reference module that matches proper names in a document
- Improves NE results by assigning entity type to previously unclassified names, based on relations with classified NEs
- May not reclassify already classified entities
- Classification of unknown entities is very useful for surnames which match a full name, or abbreviations, e.g. [Bonfield] will match [Sir Peter Bonfield]; [International Business Machines Ltd.] will match [IBM]

Named Entity Coreference



"It's an indication that Mike Armstrong is serious about local competition and serious about getting moving," said Anna-Marie Kovacs, an analyst for the brokerage firm Janney Montgomery Scott Inc. "AT&T and Teleport are going after the business market, which is where local companies make a lot of their money."

The deal also will give three major cable television companies, which are majority owners of Teleport, a collective 10 percent stake in AT&T.

By acquiring Teleport, AT&T can offer business customers local and long-distance
telephone service, and data and Internet access, under its own brand name. Using
Teleport's local facilities, the company also would be able to reduce the fees
it pays to local phone companies for access to local telephone customers.

"It's going to permit us to be much more cost-effective as we go for that local business," Armstrong said at a news briefing. "This has competition and growth written all over it." AT&T is paying for Teleport with its stock. Teleport shareholders will receive 0.943 AT&T shares for each of their Teleport shares, putting the deal at $ 59 a share based on AT&T's closing price yesterday of $ 62.62 1/2 a share, up $ 2.62 1/2. Teleport closed down $ 3.62 1/2 at $ 54.12 1/2 a share. The companies expect the deal, which must be approved by regulators

Coreference data

- Default
  - ☑ AT&T
  - ☑ Teleport
  - ☐ Hughes
  - ☐ 1997
  - ☐ yesterday
  - ☐ 1996
  - ☐ WorldCom Inc
  - ☑ Mike Armstror

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.

Argentina	B-LOC
played	O
with	O
Del	B-PER
Bosque	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 – other 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.

NER in various languages

- Arabic
  - TAGARAB (1998)
  - Pattern-matching engine + morphological analysis
  - Lots of morphological info (no differences in ortographic 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

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

NER in various languages

- ASAHARA & MATSMUTO (2003)
  - Character-based method
  - Support Vector Machine
  - 87.2% F-measure in the IREX (outperformed most word-based systems)
- Dutch
  - DE MEULDER et al. (2002)
    - Hybrid system
      - Gazetteers, grammars of names
      - Machine Learning Ripper algorithm

NER in various languages

- 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

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

NER in various languages

- Korean
  - CHUNG et al. (2003)
    - Rule-based model, Hidden Markov Model, boosting approach over unannotated data

- Portuguese
  - SOLORIO & LÓPEZ (2004, 2005)
    - Adapted CARRERAS et al. (2002b) spanish NER
    - Brazilian newspapers

NER in various languages

- 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
- Spanish
  - CARRERAS et al. (2002b)
    - Machine Learning, AdaBoost algorithm
    - BIO and Open Close approaches

NER in various languages

- 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

Named Entity Recognition

- Multilingual approaches
  - Goals - CUCERZAN & YAROWSKY (1999)
    - To handle basic language-specific evidences
    - To learn from small NE lists (about 100 names)
    - To process large and small texts
    - To have a good class-scalability (to allow the definition of different classes of entities, according to the language or to the purpose)
    - To learn incrementally, storing learned information for future use

Named Entity Recognition

- Multilingual approaches
  - GALLIPI (1996)
    - Machine Learning
    - English, Spanish, Portuguese
  - ECRAN (Extraction of Content: Research at Near Market)
  - REFLEX project (2005)
    - the US National Business Center

Named Entity Recognition

- Multilingual approaches
  - POIBEAU (2003)
    - Arabic, Chinese, English, French, German, Japanese, Finnish, Malagasy, Persian, Polish, Russian, Spanish and Swedish
    - UNICODE
    - Language independent architecture
    - Rule-based, machine-learning
    - Sharing of resources (dictionary, grammar rules…) for some languages
  - BOAS II (2004)
    - University of Maryland Baltimore County
    - Web-based
    - Pattern-matching
    - No large corpora

NER – other topics

- Character vs. word-based
  - JING et al. (2003)
    - Hidden Markov Model classifier
    - Character-based model better than word-based model
- NER translation
  - Cross-language Information Retrieval (CLIR), Machine Translation (MT) and Question Answering (QA)
- NER in speech
  - No punctuation, no capitalization
  - KIM & WOODLAND (2000)
    - Up to 88.58% F-measure
- NER in Web pages
  - wrappers

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

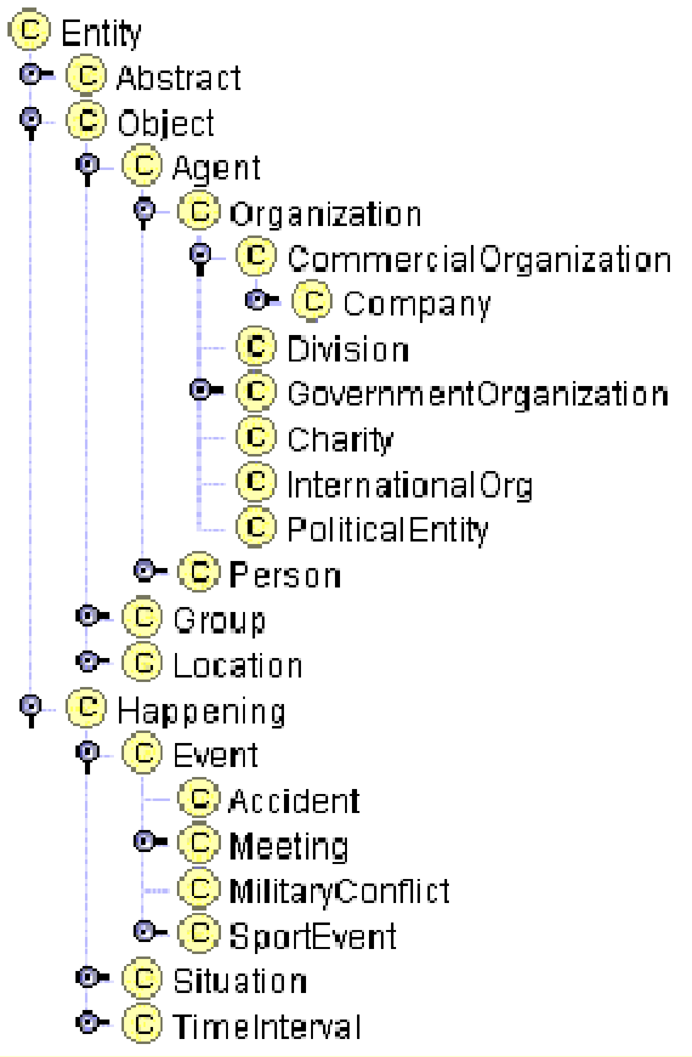
- Ekbal et al. (2012)-Hindi and Bengali
  - Active learning
  - Effective in a resource-constrained environment

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

- Need for new metrics when evaluating 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



Study Materials

- ***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.

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.

A brief introduction to Bio-text Mining

Aims: Text mining

- *Data Mining* -> needs structured data, usually in numerical form
- *Text mining*: discover & extract unstructured knowledge hidden in text–Hearst (1999)
- Text mining aids to construct hypotheses from associations derived from text
  - –protein-protein interactions
  - associations of genes–*phenotypes*
  - functional relationships among genes…etc.

An Example

- *Stress is associated with migraines*
- *Stress can lead to loss of magnesium*

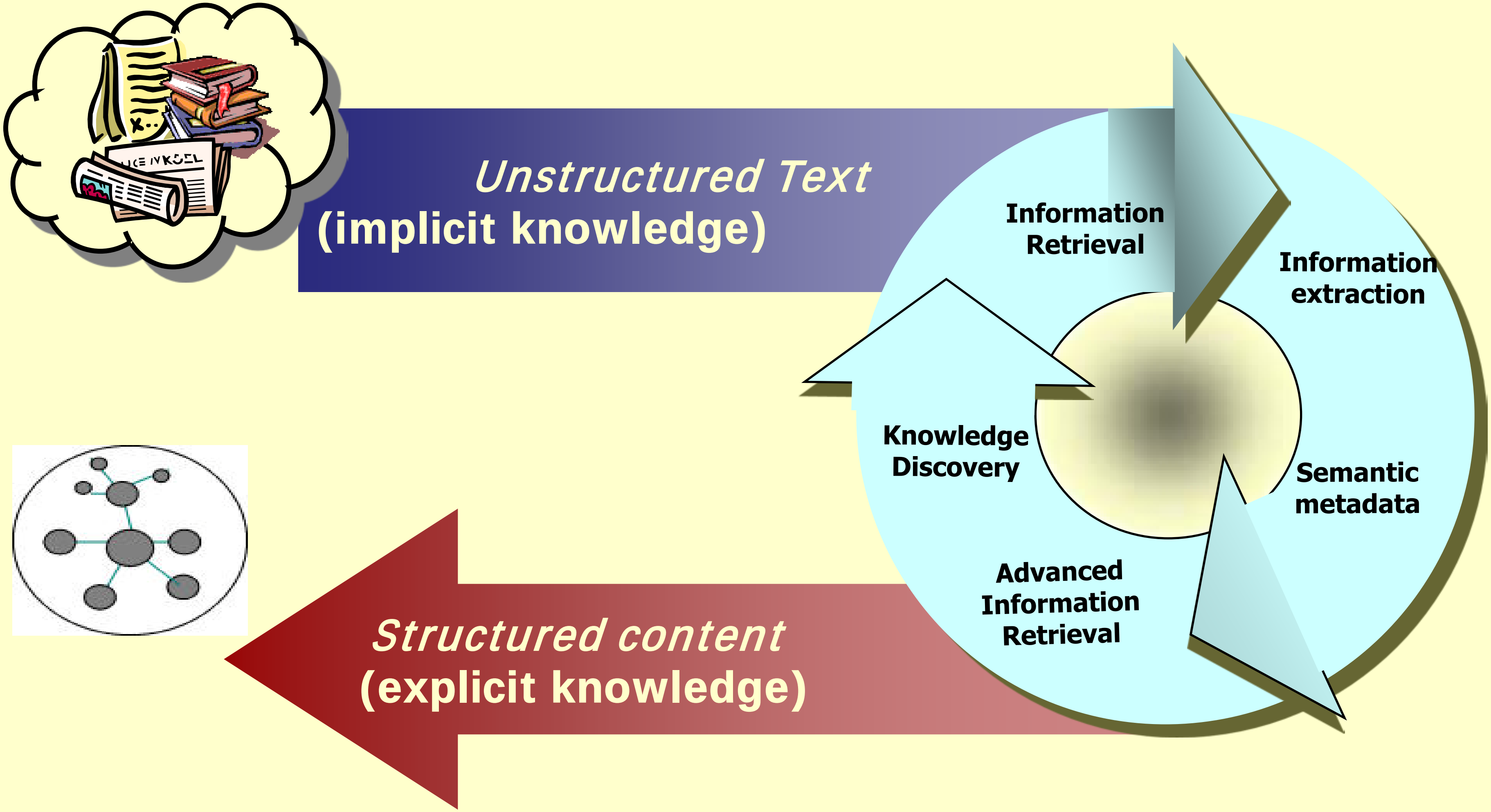
=> *Loss of magnesium may cause migraine*

Text Mining in biomedicine

- Why biomedicine?
  - Consider just MEDLINE: 23,000,000 references, 40,000-50,000 added per month
  - Dynamic nature of the domain: new terms (*genes*, *proteins*, *chemical compounds*, *drugs* etc.) constantly created
  - Impossible to manage such an information overload

From Text to Knowledge:

tackling the data deluge through text mining



Unstructured Text
(implicit knowledge)

Information Retrieval

Information extraction

Semantic metadata

Advanced Information Retrieval

Knowledge Discovery

Structured content
(explicit knowledge)

Information deluge

- *Bio-databases*, *controlled vocabularies* and *bio-ontologies* encode only small fraction of information

- Linking text to databases and ontologies
  - Curators struggling to process scientific literature
  - Discovery of facts and events crucial for gaining insights in biosciences: need for *text mining*

Impact of text mining

- Extraction of named entities (*genes, proteins, metabolites, etc.)*
- Discovery of concepts *allows* semantic annotation of documents
  - *Improves* information access by going beyond index terms, enabling semantic querying
- Construction of concept networks from text
  - *Allows* clustering, classification of documents
  - Visualisation of concept maps

Semantic annotation: An Example

- Imagine your search engine understands that "**Bangalore" is a city in “India"**, it can answer a search query on "**IT Companies in India**" with a link to a document about **Yahoo Office in Bangalore**, although the exact words "Bangalore" or "Yahoo" never occur in your search query.

Impact of TM

- Extraction of relationships (events and facts) for knowledge discovery
  - Information extraction, more sophisticated annotation of texts (*event annotation*)
  - Beyond named entities: *facts*, *events*
  - Enables even more advanced semantic querying

Challenge: the resource bottleneck

- Lack of large-scale, richly annotated corpora
  - Support training of ML algorithms
  - Development of computational grammars
  - Evaluation of text mining components

- Lack of knowledge resources: lexica, terminologies, ontologies

Some Resources for Bio-Text Mining

- Lexical / terminological resources
  - SPECIALIST lexicon, Metathesaurus (UMLS-unified medical language system)
  - Lists of terms / lexical entries (hierarchical relations)
- Ontological resources
  - Metathesaurus, Semantic Network, GO, SNOMED CT, etc
  - ➡ Encode relations among entities

Bodenreider, O. "Lexical, Terminological, and Ontological Resources for Biological Text Mining", Chapter 3, Text Mining for Biology and Biomedicine, pp.43-66

SPECIALIST lexicon

- UMLS (Unified Medical Language System) specialist lexicon http://SPECIALIST.nlm.nih.gov
  - Each lexical entry contains morphological (e.g. *cauterize, cauterizes, cauterized, cauterizing*), syntactic (e.g. complementation patterns for verbs, nouns, adjectives), orthographic information (e.g. *esophagus – oesophagus*)
  - General language lexicon with many biomedical terms (over 180,000 records)
  - Lexical programs include variation (spelling), base form, inflection, acronyms

Normalisation (lexical tools)

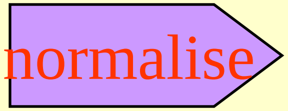
Hodgkin Disease

HODGKIN DISEASE

Hodgkin’s Disease

Hodgkin’s disease

Disease, Hodgkin ...



normalise

disease hodgkin

Steps of Norm

Remove genitive

Hodgkin's Diseases, NOS -→ Hodgkin Diseases, NOS

Replace punctuation with spaces

Hodgkin Diseases NOS

Remove stop words

Hodgkin Diseases

Lowercase

hodgkin diseases

Uninflect each word

hodgkin disease

Word order sort

disease hodgkin

Lexical tools of the UMLS
http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html

Gene Ontology

- GOA database (http://www.ebi.ac.uk/GOA/) assigns gene products to the Gene Ontology

- GO terms follow certain conventions of creation, have synonyms such as:
  - *ornithine cycle* *is an exact synonym* *of urea cycle*
  - *cell division* *is a broad synonym* *of cytokinesis*
  - *cytochrome bc1 complex* *is a related synonym* *of ubiquinol-cytochrome-c reductase activity*

GO terms, definitions and ontologies in Open Biomedical Ontologies (OBO)

id: GO:0000002

name: mitochondrial genome maintenance

namespace: biological_process

def: "The maintenance of the structure and integrity of the mitochondrial genome.“ [GOC: ai]

is_a: GO:0007005 ! mitochondrion organization and biogenesis

Metathesaurus

- Organised by concept
  - 5M names, 1M concepts, 16M relations
- built from 134 electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms
- source vocabularies
- common representation

Reading

- Book on BioTextMining
  - S. Ananiadou & J. McNaught (eds) (2006) Text Mining for Biology and Biomedicine, ArtechHouse
  - McNaught, J. & Black, W. (2006) Information Extraction, Text Mining for Biology & Biomedicine, Artechhouse, pp.143-177
- Detailed bibliography in Bio-Text Mining
  - BLIMPhttp://blimp.cs.queensu.ca/
  - http://www.ccs.neu.edu/home/futrelle/bionlp/

Bio-textmining Campaigns

Some biotext mining campaigns

- KDD Cup-2002
- TREC-Genomics (http://ir.ohsu.edu/genomics/)
- JNLPBA-2004 (http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html): Named entity recognition
- BioCreative (www.biocreative.org)-Information extraction including NER, PPI, text categorization etc. (2004, 2006, 2008,2010,2011, 2012, 2013, 2014, 2015, 2016, 2017 etc.)
- BioNLP 2009, 2011, 2013, 2015-detailed biological phenomenon

  (http://www.nactem.ac.uk/tsujii/GENIA/SharedTask

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

A Brief Introduction to Ensemble Learning

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 is 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
  - Mixtures of experts
  - Stacked generalization
  - Cascading
  - …
- Advantage: improvement in predictive accuracy

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

Why Do Ensembles Work?

Dietterich(2002) showed that ensembles overcome three problems:

- *Statistical Problem-* arises when the hypothesis space is too large for the amount of available data. Hence, there are many hypotheses with the same accuracy on the data and the learning algorithm chooses only one of them! There is a risk that the accuracy of the chosen hypothesis is low on unseen data!
- *Computational Problem-* arises when the learning algorithm cannot guarantee finding the best hypothesis.
- *Representational Problem-* arises when the hypothesis space does not contain any good approximation of the target class(es).

T.G. Dietterich, Ensemble Learning, 2002

Categories of Ensemble Learning

- Methods for Independently Constructing Ensembles
  - Bagging
  - Randomness Injection
  - Feature-Selection Ensembles
  - Error-Correcting Output Coding
- Methods for Coordinated Construction of Ensembles
  - Boosting
  - Stacking
  - Co-training

Weighted Vote based Classifier Ensemble

- Motivation
  - All classifiers are not equally good to identify all 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

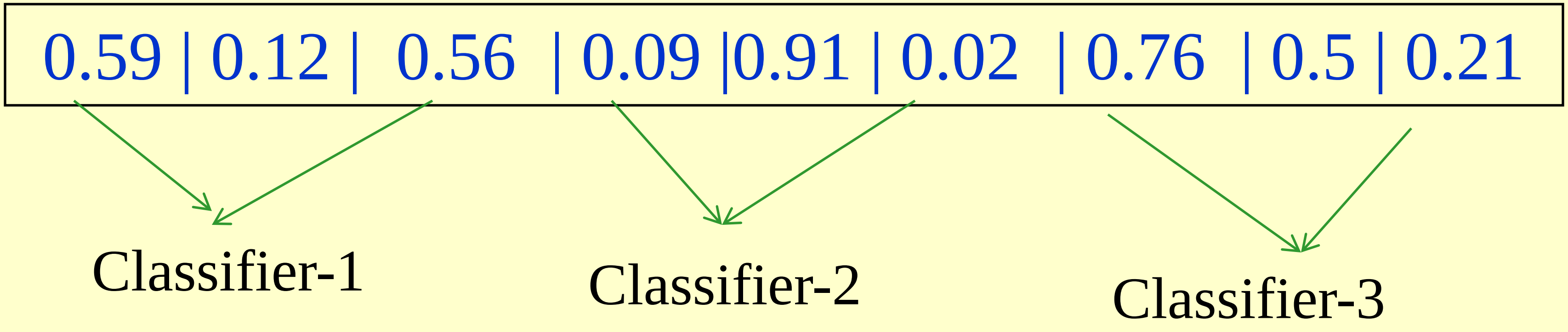
-V(i , j) ε [0, 1] denotes the degree of confidence of the *i*th classifier for the *j*th class

maximize F(B) ;

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

Here, *F1= F-measure*

Chromosome representation



0.59 | 0.12 | 0.56 | 0.09 |0.91 | 0.02 | 0.76 | 0.5 | 0.21

Classifier-1

Classifier-2

Classifier-3

- 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 test with the remaining 1/3 part

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

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

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

(b). Combined weight 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 weight selected as the joint decision

Step-4: Compute the 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 Extraction in Biomedicine

- Objective-identify biomedical entities and classify them into some predefined categories
  - *E.g. Protein, DNA, RNA, Cell_Line, Cell_Type*
- *Major Challenges*
  - building a complete dictionary for all types of biomedical NEs is infeasible due to the generative nature of NEs
  - NEs are made of very long compounded words (i.e., contain nested entities) or abbreviations and hence difficult to classify them properly
  - names do not follow any nomenclature

Challenges (Contd..)

- NEs include different symbols, common words and punctuation symbols, conjunctions, prepositions etc.
  - NE boundary identification is more difficult and challenging

- Same word or phrase can refer to different NE types based on their contexts

Features

- Context Word: Preceding and succeeding words
- Word Suffix and Prefix
    - Fixed length character strings stripped from the ending or beginning of word
- Class label: Class label(s) of the previous word (s)
- Length (binary valued): Check whether the length of the current word less than three or not (shorter words rarely NEs)
- Infrequent (binary valued): Infrequent words in the training corpus most probably NEs

Features

- Part of Speech (PoS) information- PoS of the current and/or surrounding token(s)
  - GENIA tagger V2.0.2 (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger)
- Chunk information-Chunk of the current and/or surrounding token(s)
  - GENIA tagger V2.0.2
- Unknown token feature-checks whether current token appears in training

Features

- Word normalization
  - feature attempts to reduce a word to its stem or root form (from GENIA tagger O/P)
- Head nouns
  - major noun or noun phrase of a NE that describes its function or the property
  - E.g. *factor* is the head noun for the NE *NF-kappa B transcription factor*

Features

- Verb trigger-special type of verb (e.g., *binds*, *participates* etc.) that occur preceding to NEs and provide useful information about the NE class

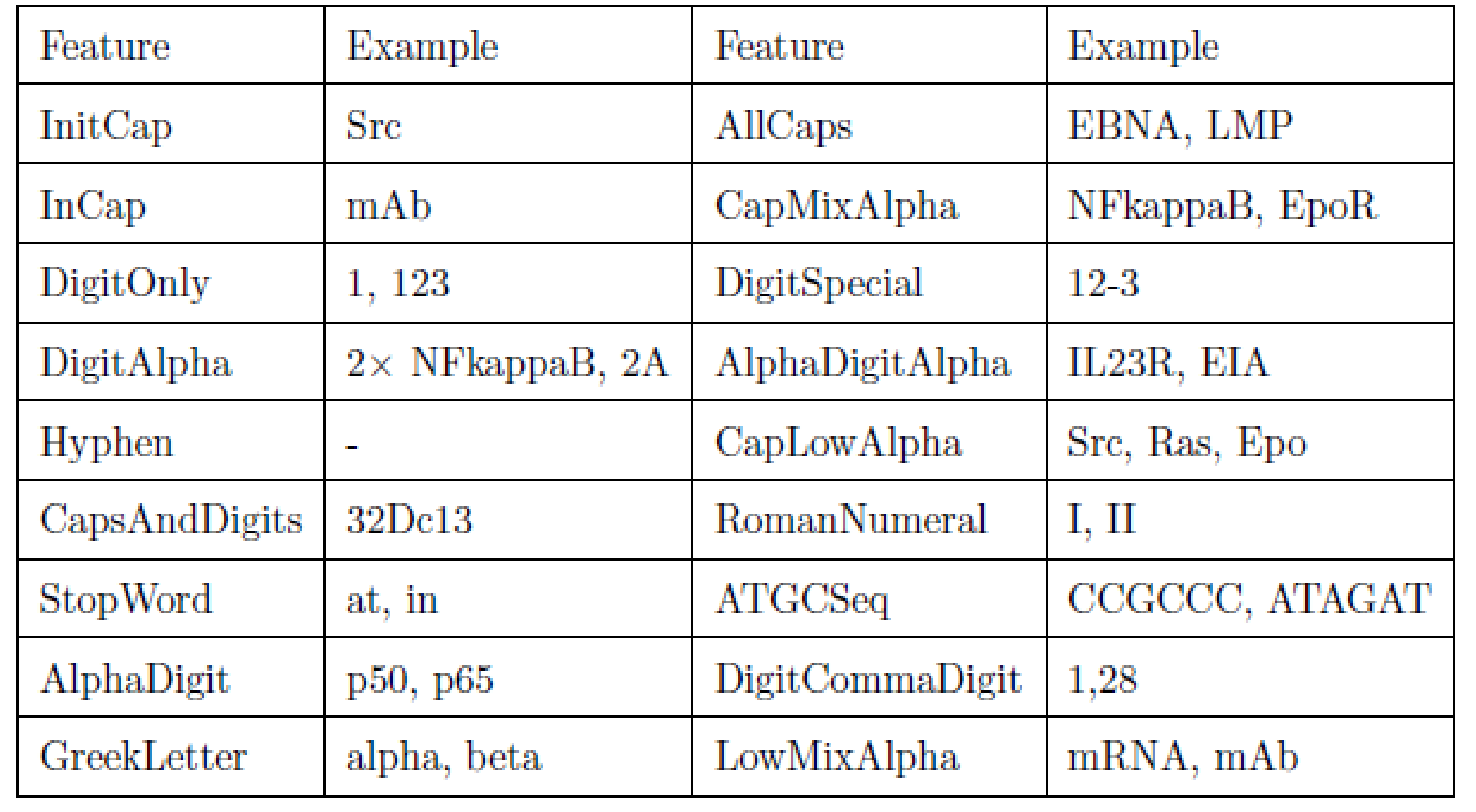
- Word class feature-Certain kinds of NEs, which belong to the same class, are similar to each other
  - capital letters→ A, small letters→a, number→O and non-English characters→-
  - consecutive same characters are squeezed into one character
  - groups similar names into the same NE class

Features

- Informative words
  - NEs are two long, complex and contain many common words that are actually not NEs
  - Function words- *of*, *and* etc.; nominals such as *active*, *normal* etc. appear in the training data often more frequently but these don't help to recognize NEs
  - Feature extracts informative words from training data statistically
- Content words in surrounding contexts-*Exploits global context information*

Features

- *Orthographic Features*-number of orthographic features depending upon the contents of the wordforms



Feature	Example	Feature	Example
InitCap	Src	AllCaps	EBNA, LMP
InCap	mAb	CapMixAlpha	NFkappaB, EpoR
DigitOnly	1, 123	DigitSpecial	12-3
DigitAlpha	2× NFkappaB, 2A	AlphaDigitAlpha	IL23R, EIA
Hyphen	-	CapLowAlpha	Src, Ras, Epo
CapsAndDigits	32Dc13	RomanNumeral	I, II
StopWord	at, in	ATGCSeq	CCGCCC, ATAGAT
AlphaDigit	p50, p65	DigitCommaDigit	1,28
GreekLetter	alpha, beta	LowMixAlpha	mRNA, mAb

Experiments

- Datasets-JNLPBA 2004 shared task datasets
  - Training: 2000 MEDLINE abstracts with 500K wordforms
  - Test: 404 abstracts with 200K wordforms
- Tagset: 5 classes
  - Protein, DNA, RNA, Cell_line, Cell_type
- Classifiers
  - CRF and SVM
- Evaluation scheme: JNLPBA 2004 shared task script (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html)
  - Recall, precision and F-measure according to *exact boundary match*, *right* and *left* boundary matching

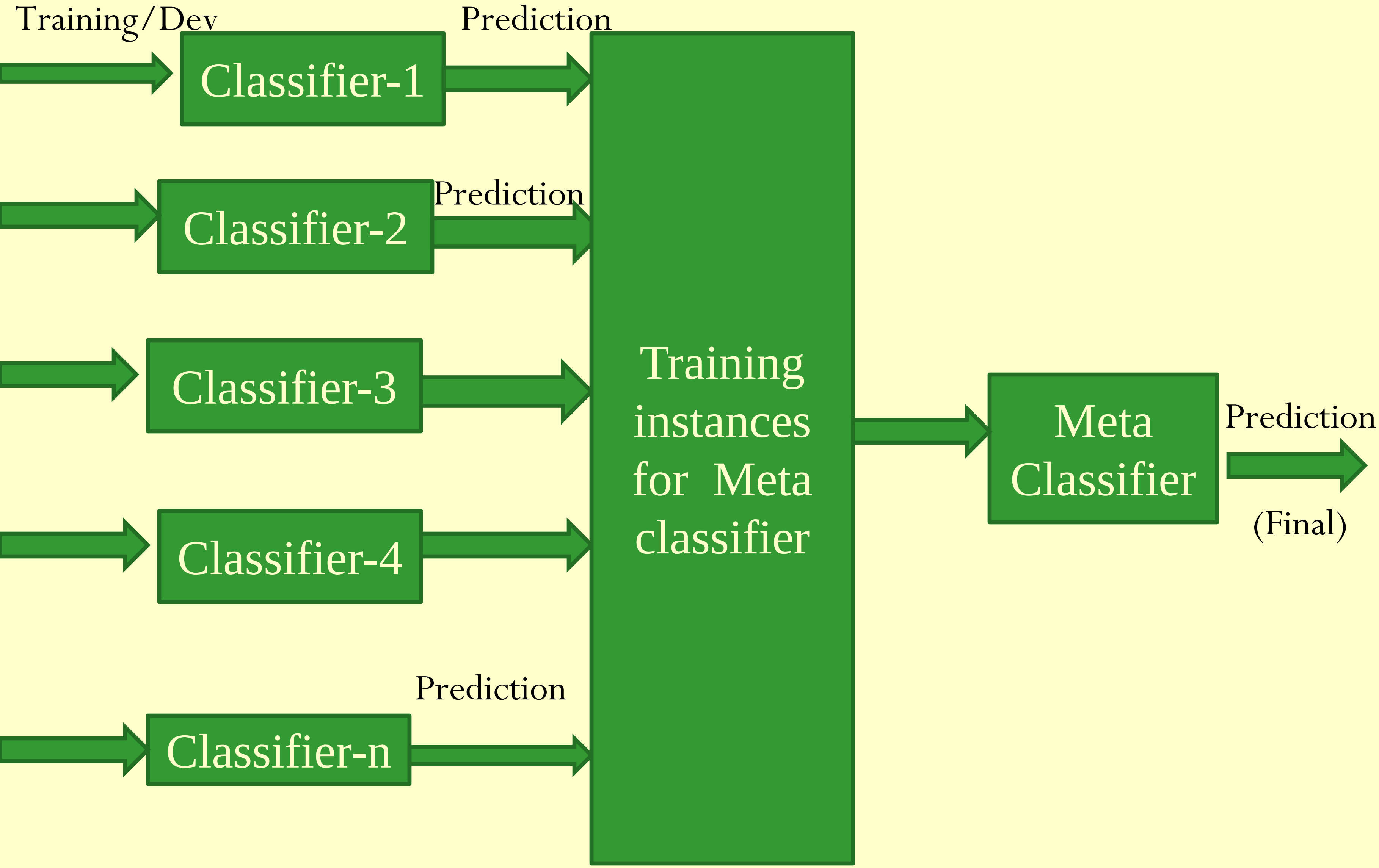
Experiments

Model	Recall	Precision	F-measure
Best individual classifier	73.10	76.76	74.76
Baseline-1	71.03	75.76	73.32
Baseline-II	71.42	75.90	73.59
Baseline-III	71.72	76.25	73.92
SOO based ensemble	74.17	77.87	75.97

- Baseline-I: Simple majority voting of the classifiers
- Baseline-II: Weighted voting where weights are based on the overall F-measure value
- Baseline-III: Weighted voting where weights are the F-measure of the individual classes

Asif Ekbal and Sriparna Saha (2013). Stacked ensemble coupled with feature selection for biomedical entity extraction, ***Knowledge Based Systems****, volume (46), PP. 22–32, Elsevier.*

Stacked Model with Feature Selection



Stacked Model with Feature Selection

- Feature selection
  - GA based
  - Select few classifiers from the final population
  - Term them as base classifiers (CRF and SVM)
- Train the base classifiers
- Evaluate on the development data
- Meta-level training instances
  - Predictions obtained on the development data
  - Original attributes

Stacked Model with Feature Selection

- For the test set
  - Generate predictions from the base classifiers
  - Use these predictions along with the original attributes as features
- Meta classifier- CRF

Experiments (JNLPBA-2004)

Model	Recall	Precision	F-measure
Best individual classifier	73.10	76.78	74.90
Majority ensemble	71.03	75.76	73.32
Weighted ensemble	71.42	75.90	73.59
Stacked ensemble	75.15	75.20	75.17

At par the state-of-the-art system

Experiments (GENETAG)

Model	Recall	Precision	F-measure
Best individual classifier	94.41	93.50	93.95
Majority ensemble	94.45	93.65	94.05
Weighted ensemble	94.67	93.91	94.29
Stacked ensemble	95.12	94.29	94.70

At par the state-of-the-art system

References

Asif Ekbal and Sriparna Saha (2013). Stacked ensemble coupled with feature selection for biomedical entity extraction, ***Knowledge Based Systems***, volume (46), PP. 22–32, Elsevier.

S.Saha, A. Ekbal and U. Sikdar (2013). Named Entity Recognition and Classification in Biomedical Text Using Classifier Ensemble. International Journal on Data Mining and Bioinformatics (in press).

A. Ekbal and S. Saha (2010). Classifier Ensemble Selection Using Genetic Algorithm for Named Entity Recognition. Research on Language and Computation (RLC), Vol. (8), PP. 73-99, Springer

A. Ekbal and S. Saha (2012). Multiobjective Optimization for Classifier Ensemble and Feature Selection: An Application to Named Entity Recognition. International Journal on Document Analysis and Recognition (IJDAR), Vol. 15(2), 143-166, Springer

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), ACM, DOI = 10.1145/1967293.1967296 http://doi.acm.org/10.1145/1967293.1967296.

Thank You for Your Attention!