Inspire…Educate…Transform.

6. Named Entity Recognition

Dr. Manish Gupta

Sr. Mentor – Academics, INSOFE

Adapted from https://gate.ac.uk/sale/talks/ne-tutorial.ppt

Course Content

- Collection of three main topics of high recent interest.
  - Search engines (Crawling, Indexing, Ranking)
    - Language Modeling
    - Text Indexing and Crawling
    - Relevance Ranking
    - Link Analysis Algorithms
  - Text Processing (NLP, NER, Sentiments)
    - Natural Language Processing
    - **Named Entity Recognition**
    - Sentiment Analysis
    - Summarization
  - Social networks (Properties, Influence Propagation, Event Detection)
    - Social Network Analysis
    - Influence Propagation in Social Networks

Agenda

- **What is NER?**

Named Entity Recognition

- NER involves
  - identification of proper names in texts, and
  - classification into a set of predefined categories of interest
- Key part of Information Extraction system
- Robust handling of proper names essential for many applications
- Pre-processing for different classification tasks

What are Named Entities?

- Person names
- Organizations (companies, government organisations, committees, etc)
- Locations (cities, countries, rivers, etc)
- Date and time expressions

- 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.

What are NOT NEs?

- Common nouns, referring to named entities – the company, the committee
- Adjectives derived from names – Bulgarian, Chinese
- Numbers which are not times, dates, percentages, and money amounts

CSE 7306c

Problems in NE Task Definition

- NE category definitions are intuitively quite clear, but there are many gray areas.
- Metonymy: substitution of the name of an attribute for that of the thing meant
- Many of these gray areas are caused by metonymy.
  - Person vs. Artefact: “The Chicken sandwich wants his bill.” vs. “Bring me a Chicken sandwich.”
  - Organisation vs. Location : “Australia won the World Cup” vs. “The World Cup took place in Australia”.
  - Company vs. Artefact: “shares in MTV” vs. “watching MTV”
  - Location vs. Organisation: “she met him at Heathrow” vs. “the Heathrow authorities”

(first cut) Solutions

- The task definition must be very clearly specified at the outset.
- Some consortia on NER essentially adopted simplistic approach of disregarding metonymous uses of words,
  - e.g. “England” was always identified as a location.
  - However, this is not always useful for practical applications of NER (e.g. football/Cricket domain).
- Idealistic solutions, on the other hand, are not always practical to implement,
  - e.g. making distinctions based on world knowledge.

Basic Problems in NE

- Variation of NEs – e.g. John Smith, Mr Smith, John.
- Ambiguity of NE types
  - John Smith (company vs. person)
  - May (person vs. month)
  - Washington (person vs. location)
  - 1945 (date vs. time)
- Ambiguity with common words, e.g. "may"

Agenda

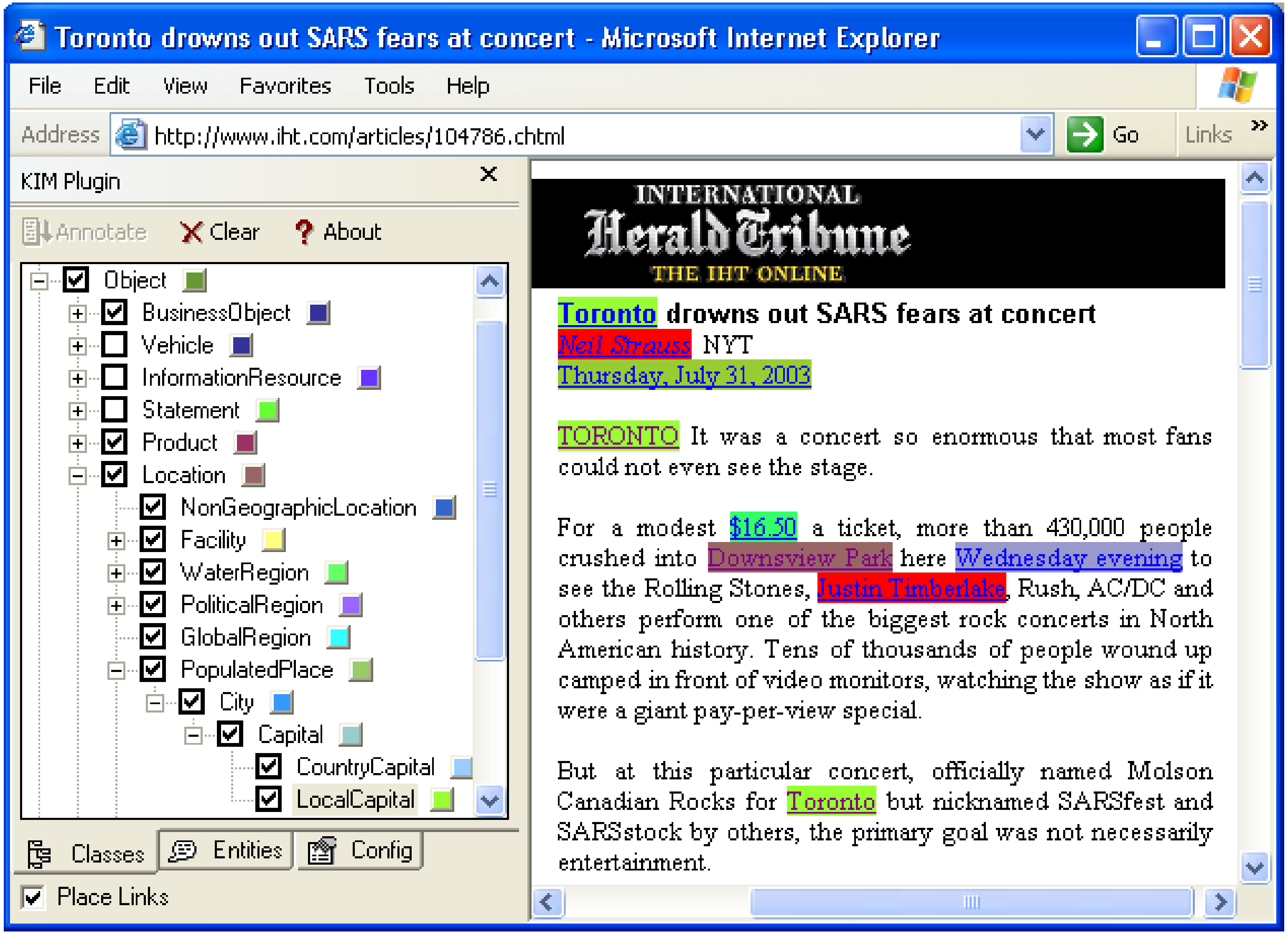
- What is NER?
- **Applications of NER**

Applications

- Can help summarisation, Speech Recognition and Machine Translation
- Question Answering
  - NER is extremely useful for systems that read text and answer queries.
  - E.g., tasks such as “Name all the colleges in Bombay listed in the document”
- Information Extraction
  - E.g., to find out and tag the subject of a web page
  - E.g., to extract the names of all the companies in a particular document
- Intelligent document access
  - Browse document collections by the entities that occur in them
  - Formulate more complex queries than IR can answer
  - Example application domains:
    - News
    - Scientific articles, e.g, MEDLINE abstracts

Application Example - KIM

Browsing by entity and ontology: http://www.ontotext.com/kim



Application Example - KIM

Search for patterns where

X, is a —Person, which name is unknown

and X involved in Y

Y, is a —Event, which name contains Kibbutz Attack

and Y took place in Z

Z, is a —Country, which name is exactly = Israel

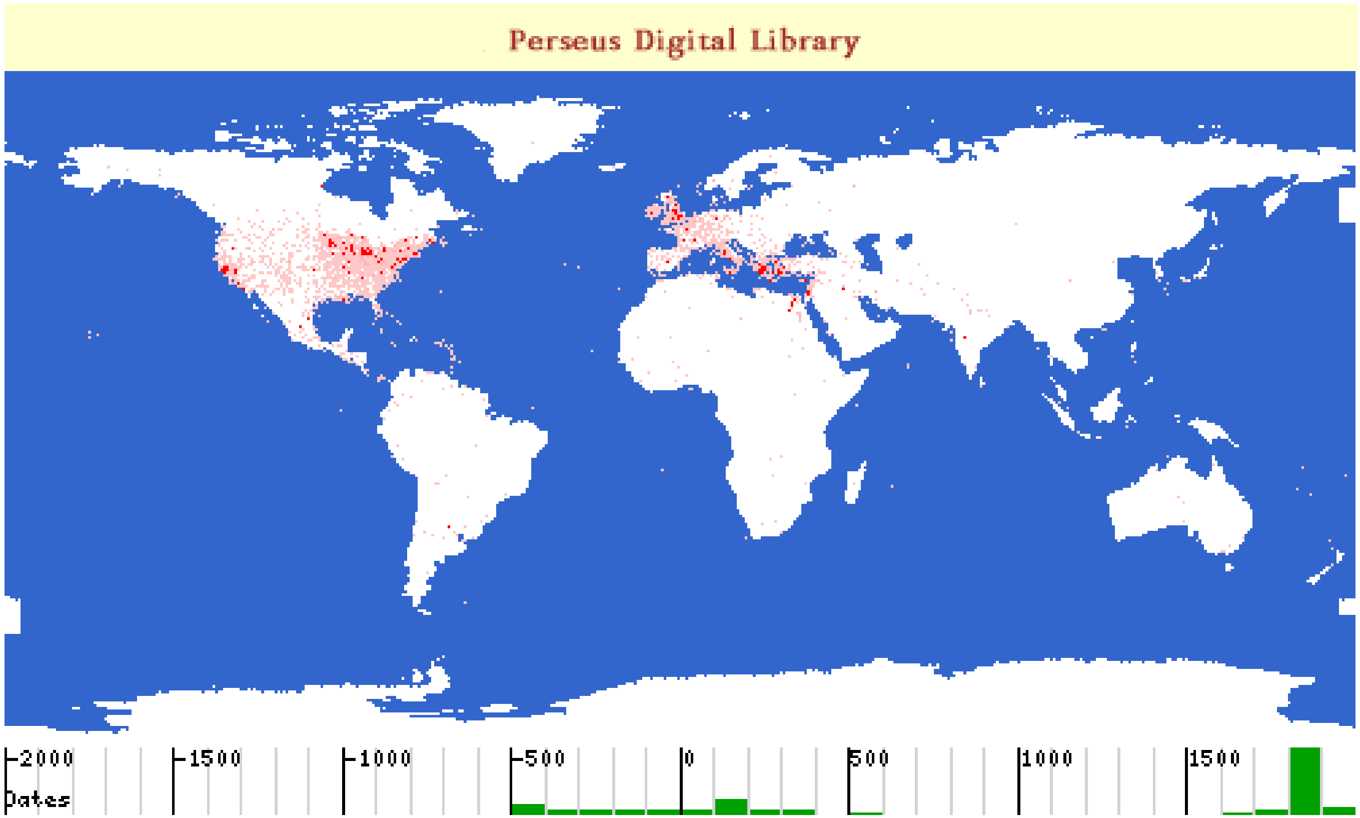
Articles (sorted by descending order)

Date & Time (GMT)	Title	Source	Origin
14/11/2002 18:57	Israel: Kibbutz attack suspect arrested [cache]	CNN	
14/11/2002 18:09	Palestinian rivals seek closer ties [cache]	BBC	
14/11/2002 17:47	Kibbutz raid suspect arrested [cache]	BBC	
14/11/2002 17:19	Israelis Capture Palestinian Suspect [cache]	FOX News	
14/11/2002 15:19	Israel Captures Militant Linked to Kibbutz Attack [cache]	Reuters	TULKARM, West Bank
14/11/2002 15:00	Palestinian killed in Nablus; kibbutz "mastermind" surrenders [cache]	AFP	
14/11/2002 13:40	Israel Troops Capture Alleged Gunman [cache]	Guardian	NABLUS, West Bank
14/11/2002 12:47	Israel Arrests Militant Linked to Kibbutz Attack [cache]	Reuters	TULKARM, West Bank
14/11/2002 11:39	Israeli army kills teenager in Nablus [cache]	AFP	
14/11/2002 08:03	Israeli tanks roll into Gaza City after army takes over Nablus [cache]	AFP	
14/11/2002 01:21	Israeli tanks enter Gaza [cache]	BBC	

1-11 of 11 Articles per page: 20

Application Example - Perseus

Time-line and geographic visualization: *http://www.perseus.tufts.edu/*



Agenda

- What is NER?
- Applications of NER
- **Evaluation and Testing**

The Evaluation Metric

- Standard metrics such as Precision, Recall, F-Measure
- We may also want to take account of partially correct answers:

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

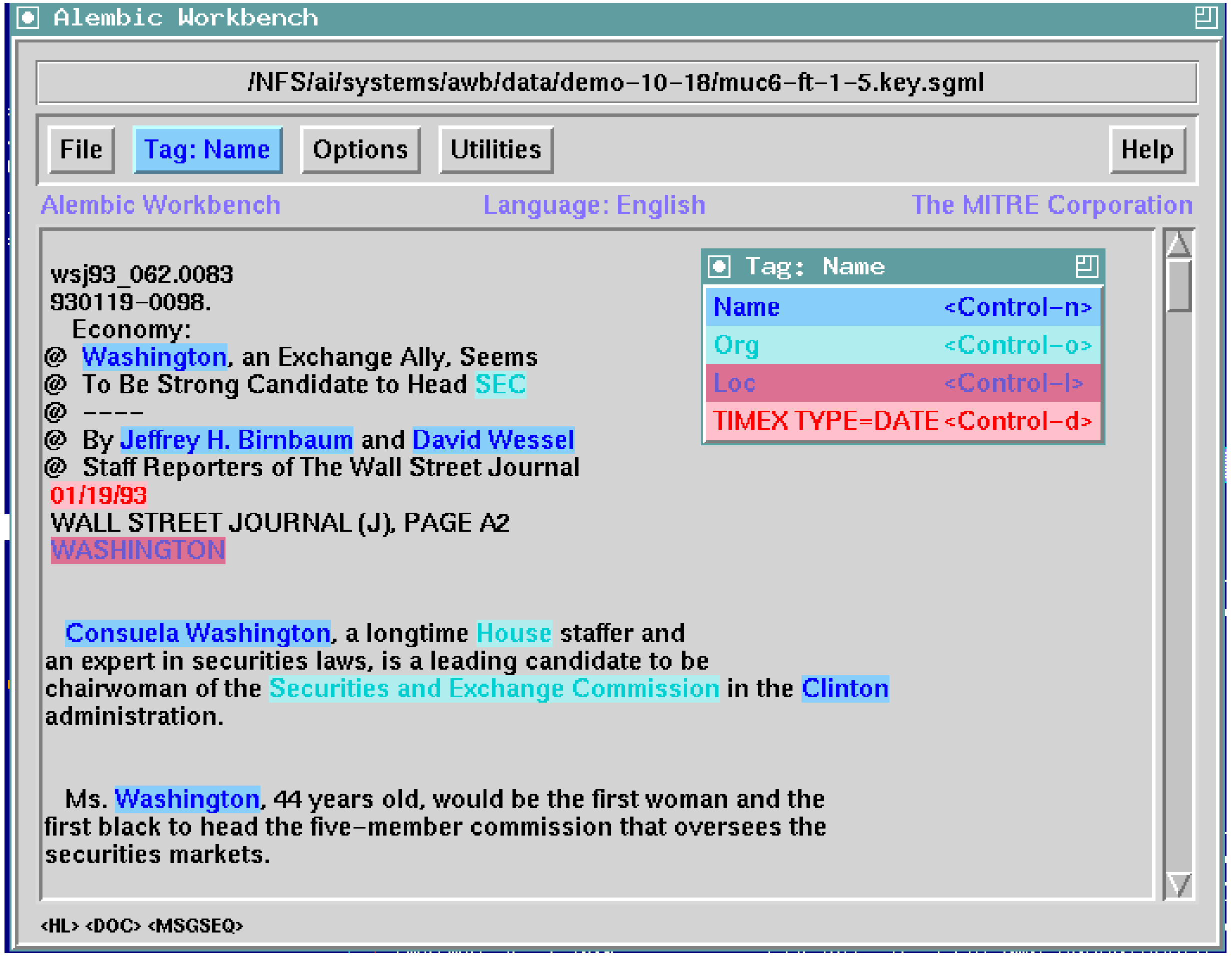
$$\text{Recall} = \frac{\text{Correct} + \tfrac{1}{2}\,\text{Partially correct}}{\text{Correct} + \text{Missing} + \text{Partial}}$$

  - Why: NE boundaries are often misplaced, so, some partially correct results!
- Scoring program – implements the metric and provides performance measures
  - For each document and over the entire corpus
  - For each type of NE

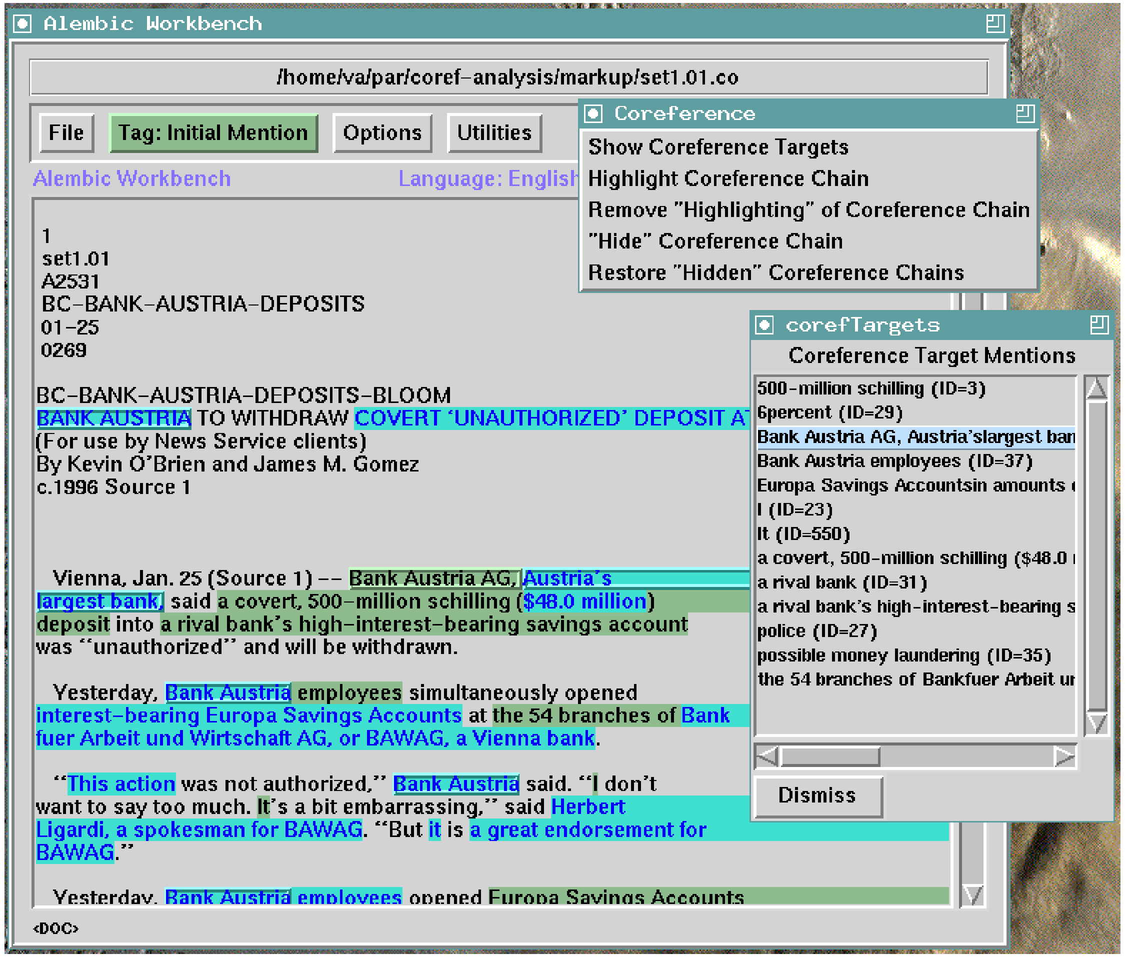
Corpora and System Development

- Corpora are divided typically into a training, validation and testing portion
- Rules/Learning algorithms are trained on the training part
- Tuned on the development/validation portion in order to optimise
  - Rule priorities, rules effectiveness, etc.
  - Parameters of the learning algorithm and the features used
- Test/Evaluation set – the best system configuration is run on this data and the system performance is obtained
- No further tuning once evaluation set is used!

NE Annotation Tools - Alembic

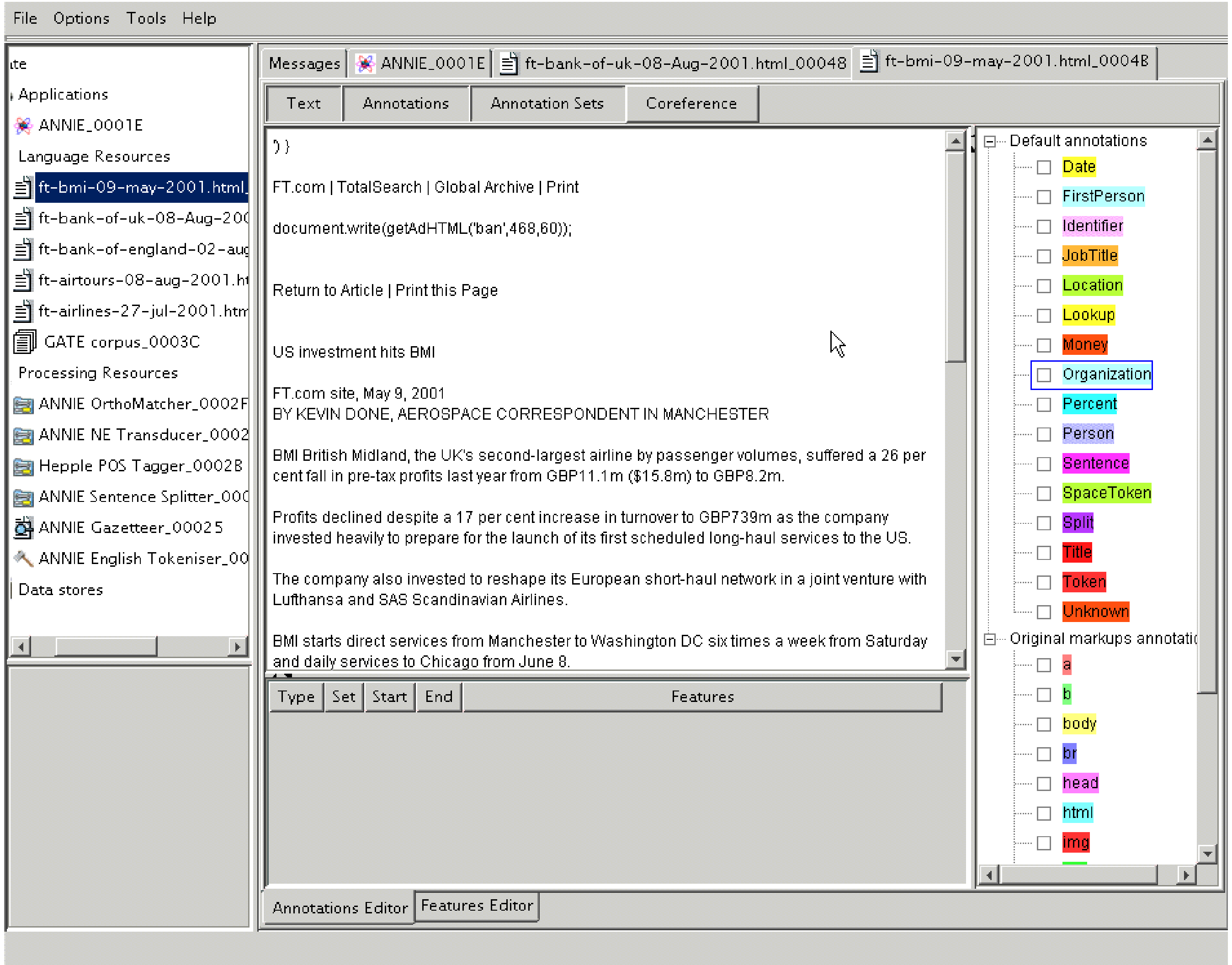


NE Annotation Tools – Alembic (2)



CSE 7306c

NE Annotation Tools - GATE



CSE 7306c

Some NE Annotated Corpora

- MUC-6 and MUC-7 corpora – English (American newswire)
  - 7 entities: Person, Location, Organization, Time, Date, Percent, Money
- CONLL shared task corpora (British newswire)
  - http://cnts.uia.ac.be/conll2003/ner/ - NEs in English and German
  - http://cnts.uia.ac.be/conll2002/ner/ - NEs in Spanish and Dutch
  - 4 entities: Person, Location, Organization, Misc
- TIDES surprise language exercise (NEs in Cebuano and Hindi)
- ACE – English - http://www.ldc.upenn.edu/Projects/ACE/
  - 7 entities: Location, Organization, Person, FAC (Facilities), GPE (Geographical/Social/Political), Vehicle, Weapon.
- BBN (Penn Treebank)
  - 12 named entity types (Person, Facility, Organization, GPE, Location, Nationality, Product, Event, Work of Art, Law, Language, and Contact-Info), nine nominal entity types (Person, Facility, Organization, GPE, Product, Plant, Animal, Substance, Disease and Game), and seven numeric types (Date, Time, Percent, Money, Quantity, Ordinal and Cardinal).
  - https://catalog.ldc.upenn.edu/LDC2005T33

Agenda

- What is NER?
- Applications of NER
- Evaluation and Testing
- **NER Approaches**

Pre-processing for NE Recognition

- Format detection
- Word segmentation (for languages like Chinese)
- Tokenisation
- Sentence splitting
- POS tagging

Two kinds of NE Approaches

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

- Statistics or machine-learning based
- Learning Systems
- developers do not need language engineering expertise
- requires large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus
- annotators are cheap

List Lookup Approach

- System that recognises only entities stored in its lists (gazetteers).
- Advantages - Simple, fast, language independent, easy to retarget (just create lists)
- Disadvantages – impossible to enumerate all names, collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Creating Gazetteer Lists

- Online phone directories and yellow pages for person and organisation names
- Locations lists
  - US GEOnet Names Server (GNS) data – 3.9 million locations with 5.37 million names
  - UN site: http://unstats.un.org/unsd/citydata
  - Global Discovery database from Europa technologies Ltd, UK
- Automatic collection from annotated training data
- Many data sources like Wikipedia, data.gov.in, linked open cloud data http://lod-cloud.net/

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, Center, River}
- e.g. Sahara City, Tyda Forest
- Cap. Word + {Street, Boulevard, Avenue, Crescent, Road}
- e.g. Preder Ghast Road, Wall Street

Problems with the Shallow Parsing Approach

- Ambiguously capitalized words (first word in sentence)
  [All Union 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]

Identification of Contextual Information

- Find windows of context around entities
- Search for repeated contextual patterns of either strings, other entities, or both
- Manually post-edit list of patterns, and incorporate useful patterns into new rules
- Repeat with new entities

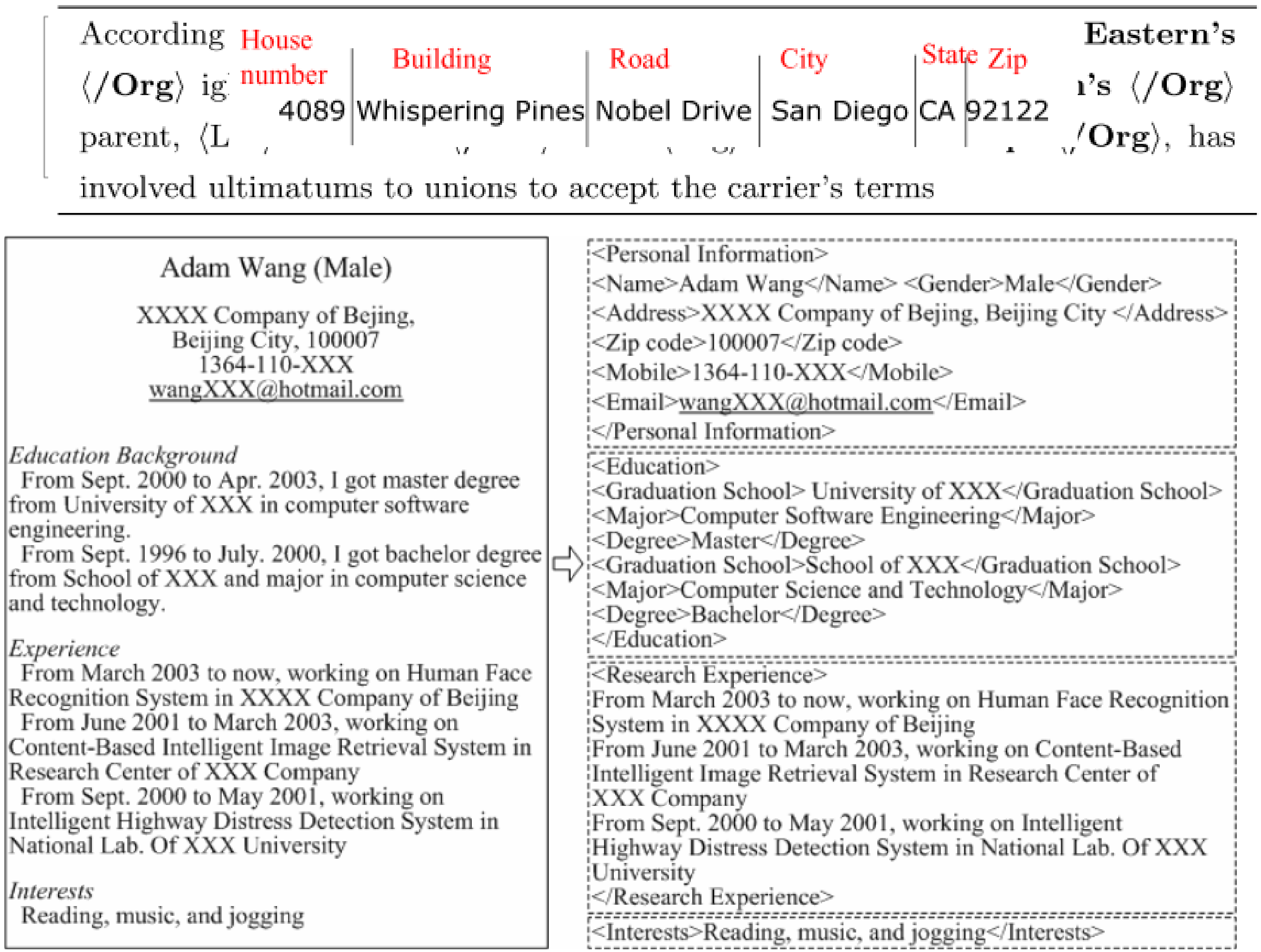
Gazetteer Lists + Rule-based NE

- Need 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

Learning Based Models

- **Machine learning approaches**
  - **Generative Models – HMM**
  - **Conditional Models – MeMM, CRF**

Why should You learn Sequence Labeling?



- Part of Speech Tagging
- Hand Writing Recognition
- Named Entity Recognition
- Structured text extraction
  - Address extraction
  - Resume extraction
- …

Machine Learning Approaches

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

India	*B-LOC*
played	*O*
with	*O*
South	*B-LOC*
Africa	*I-LOC*

Take-aways

- NER is related to identifying entities of certain types given a piece of text.
- NER is useful for many applications.
- Old NER systems were rule-based.
- New NER systems are machine learning based
  - Generative: HMMs
  - Discriminative/Conditional: MeMMs, CRFs

Further Reading

- Aberdeen J., Day D., Hirschman L., Robinson P. and Vilain M. 1995. MITRE: Description of the Alembic System Used for MUC-6. MUC-6 proceedings. Pages141-155. Columbia, Maryland. 1995.
- Black W.J., Rinaldi F., Mowatt D. Facile: Description of the NE System Used For MUC-7. Proceedings of 7th Message Understanding Conference, Fairfax, VA, 19 April - 1 May, 1998.
- Borthwick. A. A Maximum Entropy Approach to Named Entity Recognition. PhD Dissertation. 1999
- Bikel D., Schwarta R., Weischedel. R. An algorithm that learns what's in a name. Machine Learning 34, pp.211-231, 1999
- Carreras X., Màrquez L., Padró. 2002. Named Entity Extraction using AdaBoost. The 6th Conference on Natural Language Learning. 2002
- Chang J.S., Chen S. D., Zheng Y., Liu X. Z., and Ke S. J. Large-corpus-based methods for Chinese personal name recognition. Journal of Chinese Information Processing, 6(3):7-15, 1992
- Chen H.H., Ding Y.W., Tsai S.C. and Bian G.W. Description of the NTU System Used for MET2. Proceedings of 7th Message Understanding Conference, Fairfax, VA, 19 April - 1 May, 1998.
- Chinchor. N. MUC-7 Named Entity Task Definition Version 3.5. Available by from ftp.muc.saic.com/pub/MUC/MUC7-guidelines, 1997

Further reading (2)

- Collins M., Singer Y. Unsupervised models for named entity classification In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999
- Collins M. Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Perceptron. Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, pp. 489-496, July 2002 Gotoh Y., Renals S. Information extraction from broadcast news, Philosophical Transactions of the Royal Society of London, series A: Mathematical, Physical and Engineering Sciences, 2000.
- Grishman R. The NYU System for MUC-6 or Where's the Syntax? Proceedings of the MUC-6 workshop, Washington. November 1995.
- [Ign03a] C. Ignat and B. Pouliquen and A. Ribeiro and R. Steinberger. Extending and Information Extraction Tool Set to Eastern-European Languages. Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages (IESL'03). 2003.
- Krupka G. R., Hausman K. IsoQuest Inc.: Description of the NetOwlTM Extractor System as Used for MUC-7. Proceedings of 7th Message Understanding Conference, Fairfax, VA, 19 April - 1 May, 1998.
- McDonald D. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In B.Boguraev and J. Pustejovsky editors: Corpus Processing for Lexical Acquisition. Pages21-39. MIT Press. Cambridge, MA. 1996
- Mikheev A., Grover C. and Moens M. Description of the LTG System Used for MUC-7. Proceedings of 7th Message Understanding Conference, Fairfax, VA, 19 April - 1 May, 1998
- Miller S., Crystal M., et al. BBN: Description of the SIFT System as Used for MUC-7. Proceedings of 7th Message Understanding Conference, Fairfax, VA, 19 April - 1 May, 1998

Further reading (3)

- Palmer D., Day D.S. A Statistical Profile of the Named Entity Task. Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, D.C., March 31- April 3, 1997.
- Sekine S., Grishman R. and Shinou H. A decision tree method for finding and classifying names in Japanese texts. Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, Canada, 1998
- Sun J., Gao J.F., Zhang L., Zhou M., Huang C.N. Chinese Named Entity Identification Using Class-based Language Model. In proceeding of the 19th International Conference on Computational Linguistics (COLING2002), pp.967-973, 2002.
- Takeuchi K., Collier N. Use of Support Vector Machines in Extended Named Entity Recognition. The 6th Conference on Natural Language Learning. 2002
- D.Maynard, K. Bontcheva and H. Cunningham. Towards a semantic extraction of named entities. *Recent Advances in Natural Language Processing*, Bulgaria, 2003.
- M. M. Wood and S. J. Lydon and V. Tablan and D. Maynard and H. Cunningham. Using parallel texts to improve recall in IE. *Recent Advances in Natural Language Processing*, Bulgaria, 2003.
- D.Maynard, V. Tablan and H. Cunningham. NE recognition without training data on a language you don't speak. *ACL Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, Sapporo, Japan, 2003.

Further reading (4)

- H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, Y. Wilks. Multimedia Indexing through Multisource and Multilingual Information Extraction; the MUMIS project. *Data and Knowledge Engineering*, 2003.
- D. Manov and A. Kiryakov and B. Popov and K. Bontcheva and D. Maynard, H. Cunningham. Experiments with geographic knowledge for information extraction. *Workshop on Analysis of Geographic References, HLT/NAACL'03*, Canada, 2003.
- H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.
- H. Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, volume 36, pp. 223-254, 2002.
- D. Maynard, H. Cunningham, K. Bontcheva, M. Dimitrov. Adapting A Robust Multi-Genre NE System for Automatic Content Extraction. *Proc. of the 10th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2002)*, 2002.
- E. Paskaleva and G. Angelova and M.Yankova and K. Bontcheva and H. Cunningham and Y. Wilks. Slavonic Named Entities in GATE. 2003. CS-02-01.
- K. Pastra, D. Maynard, H. Cunningham, O. Hamza, Y. Wilks. How feasible is the reuse of grammars for Named Entity Recognition? *Language Resources and Evaluation Conference* (LREC'2002), 2002.

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals:	+91-9502334561/63 or 040-65743991
For Corporates:	+91-9618483483
Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering