Named Entity Recognition

Waleed butt

wb17@mars.uni-freiburg.de

Efficient Natural Language Processing

Prof. Hannah Bast

Chair of Algorithms and Data Structures

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Outline

UNI FREIBURG

- Introduction & Background
- Problem Definition
- Named Entity Recognition
- Recognizer tools & General Architecture
- Performance & Profiling
- Conclusion

Variants of Problem

- Entity Detection
  - Input: Smith is as good as his father at work.
  - 2.5Million Smiths, only in USA

- Entity Recognition
  - Input: "Dennis Ritchie was best known as the creator of the C programming language...."
  - Output:
    - PERSON http://en.wikipedia.org/wiki/Dennis_Ritchie
    - .....

Named Entity Recognition (NER)

UNI FREIBURG

- Definition:
  - "NER is the process of finding mentions of specified things in running text."

- Three universally accepted categories:
  - **Person**
    - e.g: Smith, John, Bob, Dennis
  - **Organization**
    - e.g: Google Inc, Microsoft Corporation, European Union
  - **Location.**
    - e.g Berlin, Europe, NYC

Example

Andrew Johnson was appointed as president of ACME , the biggest company in Santa Barbara, California.

[PER Andrew Johnson] was appointed as president of [ORG ACME] , the biggest company in [LOC Santa Barbara], [LOC California].

Application Areas

- Information Extraction
- Component for other areas
  - Question Answering (QA)
  - Summarization
  - Automatic translation
  - Document indexing
  - Text data mining
- Genetics & Biomedical Sciences
- Speech processing

NE Category Hierarchies

- Universilly Accepted: Person , Organization , Location
- But also:
  - Artifact, Facility, Geopolitical entity, Vehicle, Weapon, etc.
- SEKINE (2011)
  - 200 types
  - Domain-dependent
- BNN (2002)
  - 29 types
- Examples:
  - Person : Bush, Michael Jackson, Elizabeth II, LeBron
  - God : Zeus, Indra, Danu, Ra
  - Organization--> Sports_Organization:The Breen Gym, UCLA Bruins, Ma family army, Shinagawa Jogging Club

Challenges with NE Hierarchies

- Many of these grey area are caused by metonymy.
  - Washington or United states government.

- Organization vs Location
  - "England won the World Cup" vs.
  - "The World Cup took place in England".

- Location vs. Organization
  - "she met him at Heathrow" vs.
  - "the Heathrow authorities"

Message Understanding Conference

- MUC6 (1995)
  - Extraction of **Named Entities**
    - names of persons, organizations, locations
    - temporal expressions, currency and percentages
  - Tags
    - ENAMEX (“entity name expression”) tag
      - people, organization and locations
    - NUMEX (“numeric expression”) tag
      - currency and percentages
    - TIMEX (“time expression”) tag
      - temporal expressions – dates and times

NER is not ..

UNI FREIBURG

- Event recognition.
- Just matching text strings with pre-defined lists of names.
- It does not create templates, nor entity linking.

Named Entity Recognition Approaches

- List Lookup Approach
- Shallow Parsing Approach

List Lookup Approach

UNI FREIBURG

- System that recognises only entities stored in its lists. (gazetteers)
- Advantages - Simple, fast, language independent, easy to re-target
- Disadvantages – collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Shallow Parsing Approach

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

- **Location**
  - CapWord + {City, Forest, Center}
    - e.g. Sherwood Forest
  - Cap Word + {Street, Boulevard, Avenue, Road}
    - e.g. Lombard Street

Shallow Parsing Approach (cont)

UNI FREIBURG

- **External evidence** - names are often used in very predictive local contexts

- **Location**
  - "to the" COMPASS "of" CapWord
    - e.g. to the south of Freiburg
  - "based in" CapWord
    - e.g. based in Freiburg
  - CapWord "is a" (ADJ)? GeoWord
    - e.g. Freiburg is a nature friendly city

Ambiguities in Shallow Parsing

UNI FREIBURG

- Ambiguously capitalised words (first word in sentence)
  - [All American Bank] vs. All [State Police]
- Semantic ambiguity
  - "John F. Kennedy" = airport (location)
  - "Alexander Bürkle" = organization
- Structural ambiguity
  - [Cable and Wireless] vs. [Microsoft] and [Dell]
  - [Center for Computational Linguistics] vs. message from [City Hospital] for [John Smith].

Type of NER Systems

UNI FREIBURG

- Handcrafted systems
  - Knowledge (rule) based
    - Patterns
    - Gazetteers
- Automatic systems
  - Statistical
  - Machine learning
  - Unsupervised
  - Analyze: char type, POS, lexical info, dictionaries
- Hybrid systems

Named Entity Recognizer Softwares

- Stanford Named Entity Recognizer
- Illinois Named Entity Tagger
- Lingpipe Named Entity Recognizer

Stanford NER

- Working Group:
  - "The Stanford Natural Language Processing Group"
- Source code & License
  - Java + Open source (GNU GPL v2)
- Implementation
  - of linear chain CRF
- Conference
  - CoNLL03 (Person, Organization, Location).
- Feature Extraction
  - Features are more important than model

Stanford NER : Features

- Word features:
  - current word, previous word, next word, all words within a window
- Orthographic features:
  - Jenny → Xxxx
  - IL-2 → XX-#
- Prefixes and Suffixes:
  - Jenny → <J, <Je, <Jen, ..., nny>, ny>, y>
- Lots of feature conjunctions

Stanford NER: Distributed Models

UNI FREIBURG

- Trained on CoNLL, MUC and ACE
- Entities: Person, Location, Organization
- Trained on both British and American newswire, so robust across both domains
- Models with and without the distributional similarity features
-
- **Demo!**

Illinois Named Entity Tagger

UNI FREIBURG

- Java + Open source
- 90.8 F1 on CoNLL03
- External Knowledge: Wikipedia & Gazetteer list
- Non-local features
- Word Class model

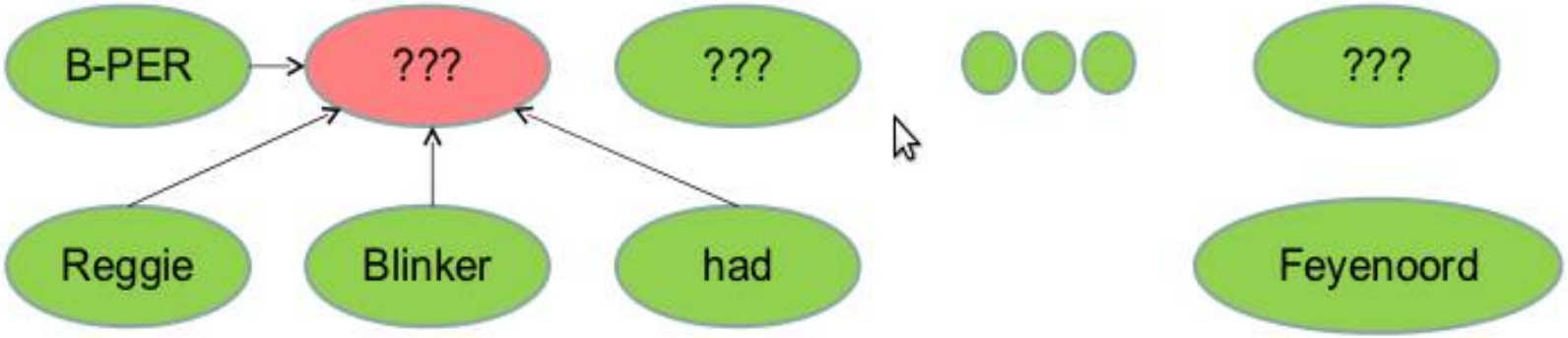
Inference & Chunk Representation

- BIO: **B**eginning **I**nside and **O**utside
- BILOU: **B**eginning, the **I**nside and the **L**ast tokens of multi-token chunks as well as **U**nit-length chunks
-

Rep. Scheme	CoNLL03 Test	CoNLL03 Dev	MUC7 Dev	MUC7 Test
BIO	89.15	**93.61**	86.76	85.15
BILOU	**90.57**	93.28	**88.09**	**85.62**

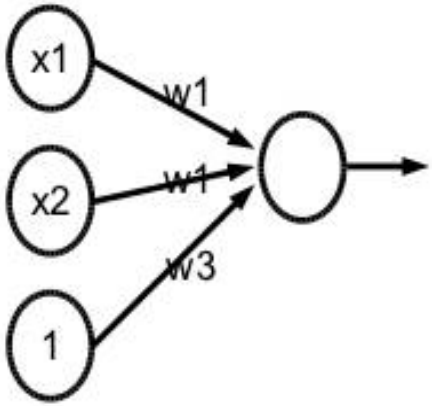
	BIO	**BILOU**
retain	O	O
the	O	O
Golan	B-loc	B-loc
Heights	I-loc	L-loc
Israel	B-loc	U-loc
captured	O	O
from	O	O
Syria	B-loc	U-loc

Modeling NER.



Use Preceptron to assign label to “Blinker” with thee following features:

- Prediction for prev word is: B-Per
- Prev word is “Reggie”
- Prev word is capitilized
- Current word is “Blinker”
- Current word is capitilized
- Next word is “had”
- ...



List of baseline features

- Tokens in the window C=[-2,+2]
- Capitalization of tokens in C.
- Previous 2 predictions
- Conjunction of previous prediction and C.
- Normalized digits (22/12/2009 ---> *DD*/*DD*/*DDDD*)
- Overall around 15 active features.

Why non-local feature?

SOCCER - BLINKER BAN LIFTED .

LONDON 1996-12-06 **Dutch]** forward **Reggie Blinker** had his indefinite suspension lifted by **FIFA** on Friday and was set to make his **Sheffield Wednesday** comeback against **Liverpool** on Saturday. **Blinker** missed his club's last two games after **FIFA** slapped a worldwide ban on him for appearing to sign contracts for both **Wednesday** and **ORG Udinese** while he was playing for **Feyenoord**.

Why non-local feature?

SOCCER - [PER BLINKER] BAN LIFTED .

[LOC LONDON] 1996-12-06 **[MISC Dutch]** forward **[PER Reggie Blinker]** had his indefinite suspension lifted by **[ORG FIFA]** on Friday and was set to make his **[ORG Sheffield Wednesday]** comeback against **[ORG Liverpool]** on Saturday. **[PER Blinker]** missed his club's last two games after **[ORG FIFA]** slapped a worldwide ban on him for appearing to sign contracts for both **[ORG Wednesday]** and **[ORG Udinese]** while he was playing for **[ORG Feyenoord]**.

External Knowledge

- Unlabeled Text
  - Word class model

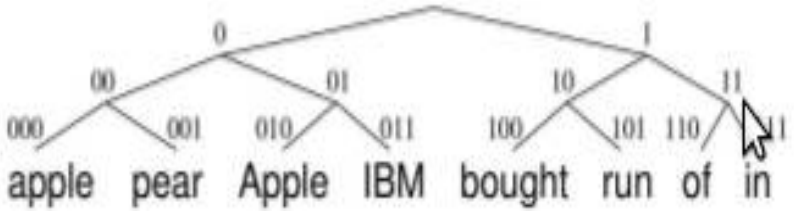


Figure 2: An extract from word cluster hierarchy.

- Gazetteers
  - 16 gazetteers, 1.5M entities from Wikipedia
  - Overall 30 gazetteers in Illinois Named Entity Tagger.
  -
  - **Demo!**

LingPipe Named Entity Recognition

UNI FREIBURG

- Commercial product (free version available)
- Java Based
- Works well for different domains. Bio, Gen & Newswire
- Regex Support
- **Demo**

Performance & Profiling

- Speed vs Accuracy
- Benchmark
  - CoNLL03 Shared task for NER
    - Reuters Corpora
      - TRC2 : comprises 1,800,370 news stories covering the period from 2008-01-01 to 2009-02-28
  - MUC7 Named Entity task
- Sample File

Results: Speeds Words per Sec

UNI FREIBURG

Input Size	Stanford	Illinois	Lingpipe
Under 100 words	626	~ 2	~50
3.3K words	1279	48	1070
37k words	1643	355	2466
3.5Mi words	Heap error	--	--
* All speeds are in **words per second**			

Speed & Memory Bootneck

UNI FREIBURG

- Stanford NER
  - Memory consumtion is biggest problem.
- Illinois Name Entity Tagger
  - Prepocessing and Gazeeter startup took huge amount of time.
    - Fast version(configuration) is available, but with less accuracy
- Lingpipe
  - Comerical version is faster and accurate.

Accuracy in term F1

UNI FREIBURG

- Illinois Named Entity Tagger
  - F1 90.8 (so best report on CoNLL03 share task)
- Stanford Entity Recognizer
  - F1 86.86 (CoNLL03)

Conclusion

- Named Entities are important in text!
- Non-local features improve the efficiency of NER.
- External Knowldge provide extra aid.
- Important sub compontnat to other part of NLP and Information extraction.

References

- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370 http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf
- Extended Named Entity Ontology with Attribute Information, Satoshi Sekine ,The Sixth International Conference on Language Resources and Evaluation ; 2008; Marrakech, Morocco
- A survey of named entity recognition and classification, David Nadeau, Satoshi Sekine, Journal of Linguisticae Investigationes 30:1 ; 2007
- Extended Named Entity Hierarchy, Satoshi Sekine, Kiyoshi Sudo and Chikashi Nobata, The Third International Conference on Language
- L. Ratinov and D. Roth, Design Challenges and Misconceptions in Named Entity Recognition, CoNLL 2009
- Reuters Corpora (RCV1, RCV2, TRC2), http://trec.nist.gov/data/reuters/reuters.html

UNI FREIBURG

Thanks

&

Questions?