Named Entity Recognition

Named Entity Recognition (NER) systems identify specific types of entities, primarily proper named entities and special categories:

- Proper Names: people, organizations, locations, etc.

  Elvis Presley, IBM, Department of the Interior, Utah

- Dates & Times: ubiquitous and surprisingly varied

  November 9, 1997, 11/9/97, 10:29 pm

- Measures: measurements with specific units

  45%, 5.3 lbs, 65 mph, $1.4 billion

- Other: Application-specific stylized terms.

  URLs, email addresses, phone numbers, social security numbers

Challenges

- No dictionary will contain all existing proper names. New names are constantly being created.
- Finding proper names isn't trivial.
  - the first word of every sentence is capitalized
  - proper names can include lower case words (e.g., UofU)
- Proper names are often abbreviated or turned into acronyms.
- But not all acronyms are proper names!
  - Ex: NLP, CS, OS

Proper Name Ambiguity

- Many companies, organizations, and locations are named after people!

  Companies: *Ford, John Hancock, Calvin Klein, Phillip Morris*

  Universities: *Brigham Young, Smith, McGill, Purdue*

  Sites: *JFK* (airport), *Washington* (capital, state, county, etc.)

- Acronyms can often refer to many different things
  - *UT, MRI, SCI* (check out the different hits on Google!)
- Many proper names can correspond to different types
  - *April, June, Georgia, Jordan, Calvin & Hobbes*

Three Common Approaches

- Hand-coded Rules
  - Good: Can perform well, esp. for specialized applications
  - Bad: Expensive to build.
- Machine Learning
  - Good: Can be easily adapted for new domains
  - Bad: Still need to annotate domain-specific texts for training
- Multilingual
  - Good: Can work across different languages
  - Bad: Usually not as effective as language-specific systems

Hand-crafted Rules

- Common prefixes and suffixes
  - *Mr., Mrs., Prof., Jr., Ph.D., Corp., Inc., Co., …*
- Lists of names, organizations, etc.
- Regular expressions for special symbols or sequences
  - dates, times, currency, urls, phone numbers, …
- Titles and Appositives
  - *President Barack Obama*
  - *the U.S. president, Barack Obama*

Message Understanding Conferences

- A series of Message Understanding Conferences (MUCs) were held in the 1990s, which played a major role in advancing information extraction research.
- MUC-3 through MUC-7 were competitive performance evaluations of IE systems built by different research groups.
- The MUCs established large-scale, realistic performance evaluations with formal evaluation criteria for the NLP community.
- The tasks included named entity recognition, event extraction, and coreference resolution.
- Some of the MUC data sets are still used as evaluation benchmarks.

A Maximum Entropy System for NER

- The MENERGI system is a nice example of a maximum entropy approach to NER:
  - *Named Entity Recognition: A Maximum Entropy Approach Using Global Information*, [Chieu & Ng, 2002]
- Good NER systems typically use a large set of *local* features based on properties of the target word and its neighboring words.
- Recent research has begun to also incorporate *global* features that capture information from across the document.

NER Types and Tagging Scheme

- MENERGI recognizes 7 types of named entities, based on the MUC-6/7 NER task definition:

  person, organization, location, date, time, money, percent

- MENERGI uses a BCEU tagging scheme:

Begin/Continue/End	*Salt*/B *Lake*/C *City*/E
Unique	*Utah*/U

- In total, the system has 29 classes:

  BCEU tags for each NE class (4x7) ⇒ 28 tags

  a special tag for Other (not a NE) ⇒ 1 tag

The Maximum Entropy Model

$$P(o \mid h) = \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(h,o)}$$

- o is the outcome (*true* or *false* with respect to a class label)
- h is the history (word)
- Z(h) is the normalization function
- $f_j(h,0)$ is a binary indicator function; k = # features
- α_j are the weights (weights are often called **parameters**; they are what the ML algorithm learns)

Example indicator function:

$$f_j(h,o) = \begin{cases} 1 & \text{if } o = \textit{true}\text{, previous word = “the”} \\ 0 & \text{otherwise} \end{cases}$$

Applying the Classifier

- One problem with NER classifiers is that they can produce inadmissible (illegal) tag sequences.

  For example, an End tag without a preceding Begin tag

- To eliminate this problem, they defined transition probabilities between classes $P(c_i \mid c_{i-1})$ to be 1 if the sequence is admissible or 0 if it is illegal.

- $P(c_i \mid s,D)$ is produced by the MaxEnt classifier.

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

Feature Set

- The classifier uses one set of *local* features, which are based on properties of the target word w, the word on its left w_{-1}, and the word on its right w_{+1}.

- The classifier also uses a set of *global* features, which are extracted from instances of the same token that occur elsewhere in the document.

- Features that occur infrequently in the training set are discarded as a form of *feature selection*.

External Dictionaries

Several external dictionaries were created by compiling lists of locations, companies, and person names.

Description	Source
Location Names	http://www.timeanddate.com http://www.cityguide.travel-guides.com http://www.worldtravelguide.net
Corporate Names	http://www.fmlx.com
Person First Names	http://www.census.gov/genealogy/names
Person Last Names	

Table 2: Sources of Dictionaries

This is very common – large lists are easy to obtain and can really help an NER system.

Summary of Local Features

- the strings of the target, previous, and next words
- the zone of the word (*headline*, *dateline*, *DD*, or main *text*)
- capitalization-based features
- is it the first word of the sentence?
- is the word in WordNet? (OOV = out-of-vocabulary feature)
- presence of the target, previous, and next words in dictionaries
- is the word a month, day, or number
- is the target word preceded/followed by an NE class prefix/suffix term
- 10 features that look for specific characters in the current word string

Target Word Character Features

Token satisfies	Example	Feature
Starts with a capital letter, ends with a period	*Mr.*	*InitCap-Period*
Contains only one capital letter	*A*	*OneCap*
All capital letters and period	*CORP.*	*AllCaps-Period*
Contains a digit	*AB3, 747*	*Contain-Digit*
Made up of 2 digits	*99*	*TwoD*
Made up of 4 digits	*1999*	*FourD*
Made up of digits and slash	*01/01*	*Digit-slash*
Contains a dollar sign	*US$20*	*Dollar*
Contains a percent sign	*20%*	*Percent*
Contains digit and period	*$US3.20*	*Digit-Period*

Table 1: Features based on the token string

Ambiguous Contexts

Some named entities occur in ambiguous contexts that can be confusing even for human readers.

McCann initiated a new global system.

The CEO of McCann announced…

The McCann family announced…

Liz Claiborne recently purchased Shoes R Us for $1.3 milllion.

She bought the shoe retailer to begin franchising it nationwide.

The company bought the shoe retailer to expand its product line.

Global Features

- Traditionally, NER systems classified each word/phrase independently of other instances of the same word/phrase in other parts of the document.
- But other contexts may provide valuable clues about what type of entity it is. For example:
  - capitalization is not indicative for the first word of a sentence
  - some contexts contain strong prefixes/suffixes in a phrase
  - some contexts contain strong preceding/following neighbors
  - acronyms can often be aligned with their expanded phrase

Summary of Global Features

- ICOC: if another occurrence of the word appears in an unambiguous position (not first word), is it capitalized?
- CSPP: do other occurrences of the word occur with a known named entity prefix/suffix?
- ACRO: if the word looks like an acronym, is there a capitalized sequence of words anywhere with these leading letters? If so, acronym features are assigned to the likely acronym word and the corresponding word sequence.
- SOIC: for capitalized word sequences, the longest substrings that appear elsewhere are assigned features.
- UNIQ: is the word capitalized and unique?

NER Results

Ablation studies look at the contribution of features or components individually to determine how much (if any) impact each one makes to the system as a whole.

	MUC-6	MUC-7
Baseline	90.75%	85.22%
+ ICOC	91.50%	86.24%
+ CSPP	92.89%	86.96%
+ ACRO	93.04%	86.99%
+ SOIC	93.25%	87.22%
+ UNIQ	93.27%	87.24%

Table 3: F-measure after successive addition of each global feature group

Summary

- Good NER systems utilize a wide variety of features, and often incorporate external dictionaries.
- Global features that look at multiple instances of a token throughout the document can improve performance.
- Hand-coded rules are expensive to create but can perform well, and can be combined with ML approaches.
- The amount of training data is important to consider when comparing results.