COM3110/4115/6115:
Text Processing

Information Extraction: Named Entity Recognition

Rob Gaizauskas

Department of Computer Science
University of Sheffield

Overview of Lectures on IE

- Introduction to Information Extraction
  - Definition + Contrast with IR
  - Example Applications
  - Overview of Tasks and Approaches
  - Evaluation + Shared Task Challenges
  - A Brief History of IE
- Named Entity Recognition
  - Task
  - Approaches to NER
  - Entity Linking
- Relation Extraction
  - Task
  - Approaches: Rule-based; Supervised learning; Bootstrapping; Distant Supervision

Named Entity Recognition: Outline

- Named Entity Recognition Task
- Approaches to NER
  - ◇ Knowledge-engineering approaches to NER
  - ◇ Supervised learning approaches to NER
- Entity Linking

Named Entity Recognition Task: Recap

- **Task**: for each textual mention of an entity of one of a fixed set of types identify its extent and its type

Cable and Wireless today announced . . .	Extent: 0-3; Type = ORG
IBM and Microsoft today announced . . .	Extent: 0-1; Type = ORG
	Extent: 2-3 Type = ORG
John Lewis hired . . .	Extent: 0-2; Type = ORG
Theresa May hired . . .	Extent: 0-2; Type = PER

- Types of entities which have been addressed by IE systems include:
  - ◇ Named individuals
    - Organisations, persons, locations, books, films, ships, restaurants . . .
  - ◇ Named Kinds
    - Proteins, chemical compounds/drugs, diseases, aircraft components . . .
  - ◇ Times
    - temporal expressions dates, times of day
  - ◇ Measures
    - monetary expressions, distances/sizes, weights . . .

Entity Extraction – Coreference: Recap

- Multiple references to the same entity in a text are rarely made using the same string:
  - Pronouns – Tony Blair . . . he
  - Names/definite descriptions – Tony Blair . . . the Prime Minister
  - Abbreviated forms – Theresa May ... May; United Nations ... UN
  - Orthographic variants – alpha helix . . . alpha-helix . . . α-helix . . . a-helix
- Different textual expressions that refer to the same real world entity are said to corefer.
- Clearly IE systems are more useful if they can recognise which text mentions are coreferential.
- **Coreference Task**: link togther all textual references to the same real world entity, regardless of whether the surface form is a name or not
- Detecting which entity mentions corefer may or may not be treated as a task separate to that of recognising entity mentions

Named Entity Recognition: Outline

- Named Entity Recognition Task
- Approaches
  - Knowledge-engineering approaches to NER
  - Supervised learning approaches to NER
- Entity Linking

Overview of Approaches to NER

As with IE in general approaches to NER may be placed into four categories:

- Knowledge Engineering Approaches
- Supervised Learning Approaches
- Bootstrapping Approaches
- Distant Supervision Approaches

For reasons of time, we will consider the first two only.

Knowledge Engineering Approaches to NER

- Such systems typically use
  - named entity lexicons and
  - manually authored pattern/action rules or regular expression/FST recognisers
- Dominant approach in the 1990s and still in use in many IE systems today.
- One such NER system, developed for participation in MUC-6, is described in Wakao et al. (1996) – will use as an example.
- The Wakao et al. system recognizes organisation, person and location names and time expressions in newswire texts
- System has three main stages:
  - Lexical processing
  - NE parsing
  - Discourse interpretation

Knowledge Engineering Approaches to NER
Step 1: Lexical Processing

- Many rule-based NER systems made extensive use of specialised lexicons of proper names, such as gazetteers – lists of place names
- The Wakao et al. system has specialised lexicons for
  - ◇ Organisations (2600 entries)
  - ◇ Locations (2200 entries)
  - ◇ Person names (500 entries)
  - ◇ Company designators (e.g. Plc,Corp, Ltd – 94 entries)
  - ◇ Person titles (e.g. Mr, Dr, Reverend – 160 titles)
- Why not use even larger gazetteers?
  - ◇ e.g. Gazetteer of British Place Names claims it "provides an exhaustive Place Name Index to Great Britain, containing over 50,000 entries"
- Reasons:
  - ◇ Many NEs occur in multiple categories – the larger the lexicons the greater ambiguity, e.g.,
    - Ford – company vs Ford – person vs Ford – place
  - ◇ the listing of names is never complete, so need some mechanism to type unseen NEs in any case

Knowledge Engineering Approaches to NER
Step 1: Lexical Processing (cont)

Principal lexical processing sub-steps in the Wakao et al. system are:

- Tokenisation, sentence splitting, morphological analysis
- Part-of-speech tagging – tags known proper name words and unknown uppercase-initial words as proper names (NNP, NNPS)
- Name List/Gazetter Lookup and Tagging (organisations, locations, persons, company designators, person titles)
- Trigger Word Tagging – certain words in multi-word names function as trigger words, permitting classification of the name
  - ◇ e.g. Airlines in Wing and Prayer Airlines
  - ◇ system has trigger words for various orgs, govt institutions, locations

Example:
Norwich Investment Bank plc. today announced ... $\longrightarrow$
Norwich$_{NNP/LOC}$ Investment$_{NNP}$ Bank$_{NNP/ORG-TRIGGER}$ plc.$_{NN/CDG}$ today$_{RB}$ announced$_{VBD}$ ...

Knowledge Engineering Approaches to NER
Step 2: NE Parsing

- After lexical processing the next step in the Wakao et al. system is NE parsing.
- The system has 177 hand-produced rules for proper names: 94 for organisation; 54 for person; 11 for location; 18 for time expressions.
- A fragment of the proper name grammar:

```
NP--> ORGAN_NP
ORGAN_NP --> LIST_LOC_NP NAMES_NP CDG_NP
ORGAN_NP --> LIST_ORGAN_NP NAMES_NP CDG_NP
ORGAN_NP --> NAMES_NP ‘& NAMES_NP
NAMES_NP --> NNP NAMES_NP
NAMES_NP --> NNP
```

- The rule `ORGAN NP --> NAMES_NP ‘&’’ NAMES_NP` means:

  If an unclassified proper name (`NAMES_NP`) is followed by ‘& and another unclassified proper name, then it is an organisation name.

  E.g. Marks & Spencer and American Telephone & Telegraph

Knowledge Engineering Approaches to NER
Step 3: Discourse Interpretation – Coreference Resolution

- When the name class of an antecedent (anaphor) is known then establishing coreference allows the name class of the anaphor (antecedent) to be established.
- An unclassified PN may be co-referential with a variant form of a classified PN, e .g.
  - ◇ Ford – Ford Motor Co.
  - ◇ CAA – Creative Artists Agency

  In such cases the unclassified PN may be inferred to have the same class as the classified PN.
  Wakao et al. use 45 heuristics of this type for organisation, location, and person names.
- An unclassified PN may be co-referential with a definite NP which permits the PNs class to be inferred
  - ◇ E.g. Kellogg ... the breakfast cereal manufacturer

Knowledge Engineering Approaches to NER
Step 3: Discourse Interpretation – Semantic Type Inference

Semantic type information about the arguments in certain syntactic relations is used to make inferences permitting the classification of PNs:

- noun-noun qualification: when an unclassified PN qualifies an organisation-related object then the PN is classified as an organisation; e.g. Erickson stocks
- possessives: when an unclassified PN stands in a possessive relation to an organisation post, then the PN is classified as an organisation; e.g. vice president of ABC, ABCs vice president
- apposition: when an unclassified PN is apposed with a known orgnisation post, the former name is classified as a person name; e.g. Miodrag Jones, president of XYZ
- verbal arguments: when an unclassified PN names an entity playing a role in a verbal frame where the semantic type of the argument position is known, then the name is classified accordingly; e.g. Smith retired from his position as . . . (subject type of retire is PERSON)

Knowledge Engineering Approaches to NER: Evaluation of Wakao et al.

- Evaluated on MUC-6 NE evaluation set – a blind test set of 30 Wall Street Journal Articles containing:
  - 449 organisation names
  - 373 person names
  - 110 location names
  - 111 time expressions
- Results were:

Proper Name Class	Recall	Precision
Organisation	91%	91%
Person	90%	95%
Location	88%	89%
Time	94%	97%
Overall	91%	93%

- Best system results on this evaluation were F-measure = 96.42%
  - Human results were 96.68%

Knowledge Engineering Approaches to NER: Strengths and Weaknesses

Strengths

- High performance – only several points behind human annotators
- Transparent – easy to understand what system is doing/why

Weaknesses

- Porting to another domain requires substantial rule re-engineering
- Acquisition of domain-specific lexicons
- Rule writing requires high levels of expertise

Named Entity Recognition: Outline

- Named Entity Recognition Task
- Approaches
  - ◇ Knowledge-engineering approaches to NER
  - ◇ Supervised learning approaches to NER
- Entity Linking

Supervised learning approaches to NER

- Supervised learning approaches aim to address the portability problems inherent in knowledge engineering NER
  - Instead of manually authoring rules, systems learn from annotated examples
  - Moving to new domain requires only annotated data in the domain – can be supplied by domain expert without need for expert computational linguist
- A wide variety of supervised learning techniques have been tried, including
  - Hidden Markov models
  - Decision Trees
  - Maximum Entropy
  - Support Vector Machines
  - Conditional Random Fields
  - AdaBoost
  - Deep Learning

Supervised learning approaches to NER: Sequence Labelling

- Systems may learn
  - patterns that match extraction targets
  - classifiers that label tokens as beginning/inside/outside a tag type
- Most work in recent years has followed the latter approach – called sequence labelling.
- In sequence labelling for NER, each token is given one of three label types:
  - B_{Type} if the token is at the beginning of a named entity of type $= Type$ (here, e.g., $Type \in \{ORG, PER, LOC\}$).
  - I_{Type} if the token is inside a named entity of type $= Type$
  - O if the token is outside any named entity

  For obvious reasons this scheme is called BIO or sometimes IOB sequence labelling

Supervised learning approaches to NER: Sequence Labelling – Example

- Suppose we have the sentence

  [$_{ORG}$ American Airlines], a unit of [$_{ORG}$ AMR Corp.], immediately matched the move, spokesman [$_{PER}$ Tim Wagner] said.

  (Jurafsky and Martin, 2nd ed., p. 730)

- In BIO encoding this example looks like this ⟶
- Given labelled sequences like this example sentence as training data, the task of the supervised learner to learn to predict the labelling of a new, unlabelled example.

Words	**Label**
American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PERS}
Wagner	I_{PERS}
said	O
.	O

Supervised learning approaches to NER: Features for Sequence Labelling

- Given a BIO-type encoding, each training instance (token) is typically represented as a set of features.
- Features can be not only characteristics of the token itself but of neighbouring tokens as well
  - usually consider tokens in a window of e.g. $\pm$ 2 or3 tokens either side of the training instance
- Features commonly used for NER sequence labelling include:

Feature	**Explanation**
Lexical items	The token to be labeled
Stemmed lexical items	Stemmed version of the target token
Shape	The orthographic pattern of the target word
Character affixes	Character-level affixes of the target and surrounding words
Part of speech	Part of speech of the word
Syntactic chunk labels	Base-phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of words/Bag of N-grams	Words and/or N-grams occurring in the surrounding context

(Jurafsky and Martin, 2nd ed., p. 731)

Supervised learning approaches to NER: Features for Sequence Labelling (cont)

- For case sensitive languages like English the orthographic pattern of a token carries significant information.
- Commonly used "shape" features include:

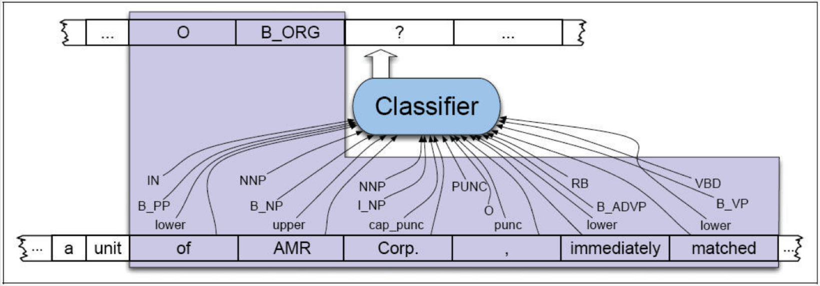
Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character with period	H.
Ends in digit	A9
Contains hyphen	H-P

(Jurafsky and Martin, 2nd ed., p. 731)

Supervised learning approaches to NER: Features for Sequence Labelling (cont)

- After a model has been learned, then at classification time the classifier extracts features from
  - ◇ the input string
  - ◇ its left predictions

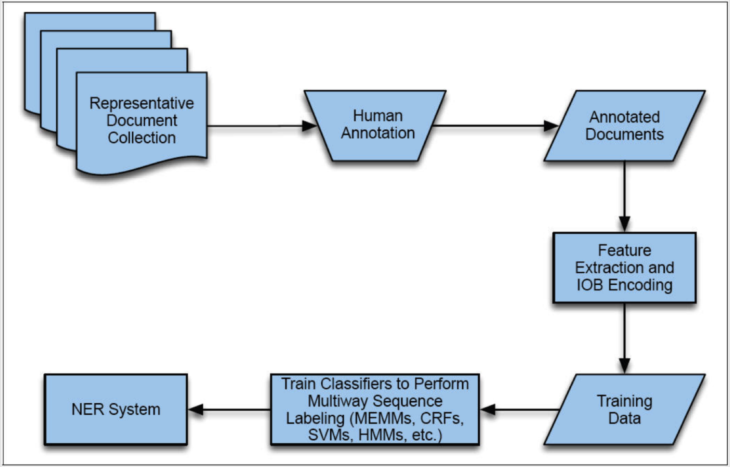
  The available features for classification are those shown in the shaded area in the following figure:



(Jurafsky and Martin, 2nd ed., p. 733)

Supervised learning approaches to NER: Sequence Labelling Overview

- The following diagram recaps the main steps in the sequence labelling approach to NER.



Supervised learning approaches to NER: Carreras et al. (2003)

- One implementation of the BIO-based sequence labelling for NER is described by Carreras et al. (2003)
  - ◇ Achieved highest score in the CONLL 2003 NER shared task challenge.
- Notable aspects of their approach include:
  - ◇ They divided the problem into two parts
    - NE detection: in a first pass over the text BIO tags are assigned without regard to type – i.e. boundaries are found for all NE's regardless of whether they are organisations, persons, locations, etc.
    - NE classification: in a second pass the NE's detected in the first pass are assigned a class (organisation, person, location, etc.)
    - Two pass approach has the advantage that training data for all NE classes can be used for the NE detection task
  - ◇ They used the Adaboost classifier
  - ◇ They used all features mentioned above plus some additional ones, e.g.
    - Type pattern of consecutive words in context – functional (f), capitalized (C), lowercased (l), punctuation mark (.), quote (), other (x) – e.g. word type pattern for the phrase John Smith payed 3 euros is CClxl.

Supervised learning approaches to NER: Carreras et al. (2003)

- Overall performance on NE Detection:
  - 91.93% precision/94.02% recall on English test set
  - 85.85% precision/72.61% recall on German test set
    - Note: all common nouns are capitalised in German
- Overall best performance on NE Classification, assuming perfect Detection:
  - 95.14% accuracy for English
  - 85.14% accuracy for German
- Overall performance for NE Detection + Classification:
  - 84.05% precision/85.96% recall on English test set
  - 75.47% precision/63.82% recall on German test set
- Looking at different entity classes, LOC and PER score consistently higher than ORG and MISC

Named Entity Recognition: Outline

- Named Entity Recognition Task
- Approaches
  - Knowledge-engineering approaches to NER
  - Supervised learning approaches to NER
- Entity Linking

Entity Linking

- One important application of IE is knowlege base population (KBP) – facts are gathered from open access web sources and used to build a structured information repository.
- For KBP to work, not only must entities be detected, they must be linked to the appropriate entry in the KB, if facts are to be correctly assembled.
- This leads to the Entity Linking Task: Given a text with a recognised NE mention in that text and a knowledge base (KB), such as Wikipedia, link the NEs to the matching entry in the KB if there is one, else create an entry.
- Is this task difficult? – yes!!
  - ◇ Wikipedia contains over 200 entries for John Smith
  - ◇ There are at least 1,716 places called San José (or San Jose); 41 Springfield's in the US
  - ◇ Ashoka Restaurant, ABC Taxis, . . .

Entity Linking (cont)

- Many approaches have been developed.
- Simple approach: given a text T containing an NE mention m and using Wikipedia as a KB
  1. index all pages in the KB using an information retrieval system
  2. build a query from T (e.g. use the sentence/paragraph/whole text) containing m and search the KB
  3. from the ranked list of KB pages returned by step 2 pick the high ranked page whose name matches m and return it

  Problem: doesm't work very well
- More successful approaches consider disambiguating all NEs jointly
  - ◇ Intuition: in disambiguating a text mentioning Ashoka and Sheffield, the Ashoka mentioned is likely to be in Sheffield, while the Sheffield is likely to be one containing an Ashoka restaurant.
  - ◇ See, e.g., Alhelbawy and Gaizauskas (2014)

Conclusion

- Named Entity Recognition (NER) is a core IE technology that is now relatively mature and at "usable" performance levels
- NER aims to detect and classify all mentions of named entities of a given set of entity types within a given text
- Techniques used have included:
  - ◇ knowledge engineering approaches
  - ◇ supervised learning approaches
    - a common approach here is to use BIO sequence labelling
- Open challenges include:
  - ◇ reducing the amount of training data needed via, e.g. bootstrapping techniques
  - ◇ exploiting existing structured data sources to generate "weakly labelled" training data (aka distant supervision)
  - ◇ expanding the classes of entities addressed
  - ◇ developing NERs for languages other than English

References

- Alhelbawy, A. and Gaizauskas, R. Collective Named Entity Disambiguation using Graph Ranking and Clique Partitioning Approaches. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 1544–1555, 2014.
- Carreras, X., Marquez, L. and Padro, L. A Simple Named Entity Extractor using AdaBoost. Proceedings of CoNLL-2003 , 152–155, 2003.
- Jurafsky, D and Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Comp2nutational Linguistics and Speech Recognition. 2nd ed. Pearson Inc. 2009. See Chapter 22.1 "Named Entity Recognition".
- Wakao, T., Gaizauskas, R. and Wilks, Y. Evaluation of an Algorithm for the Recognition and Classification of Proper Names. In Proceedings of the 16th International Conference on Computational Linguistics (COLING96), 418-423, 1996.