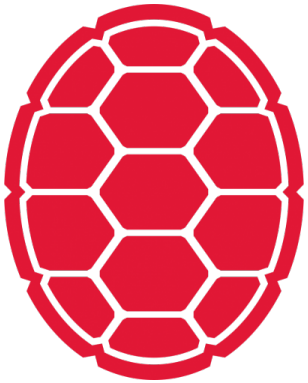

Entity-Driven Desiderata

Computational Linguistics: Jordan Boyd-Graber
University of Maryland

NAMED ENTITY RECOGNITION

Adapted from slides by Luke Zettlemoyer

Information Extraction

- IE = extracting information from text
- Sometimes called text analytics commercially
- Extract entities
  - People, organizations, locations, times, dates, prices, . . .
  - Or sometimes: genes, proteins, diseases, medicines, . . .
- Extract the relations between entities
  - Located in, employed by, part of, married to, . . .
- Figure out the larger events that are taking place

Named Entity Recognition

The [European Commission ORG] said on Thursday it disagreed with [German MISC] advice.
Only [France LOC] and [Britain LOC] backed [Fischler PER] 's proposal .

"What we have to be extremely careful of is how other countries are going to take [Germany LOC] 's lead", [Welsh National Farmers ' Union ORG] ([NFU ORG]) chairman [John Lloyd Jones PER] said on [BBC ORG] radio .

Classify them by type, usually {ORG, PER, LOC, MISC}

Named Entity Recognition (NER)

- ItâĂŹs a tagging task, similar to part-of speech (POS) tagging
- So, systems use sequence classifiers: HMMs, MEMMs, CRFs
- Features usually include words, POS tags, word shapes, orthographic features, gazetteers, etc.
- Accuracies of >90% are typical âĂŤ but depends on genre!
- NER is commonly thought of as a âĂİsolved problemâĂİ
- A building block technology for relation extraction
- E.g., http://nlp.stanford.edu/software/CRF-NER.shtml

Tagging for NER

BIO = Begin, Inside, Outside

Relations from Resources

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Learning Relations from Text

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...

Bootstrapping Relations

- Requires that we have seeds for each relation
  - Sensitive to original set of seeds
- Big problem of semantic drift at each iteration
- Precision tends to be not that high
- Generally have lots of parameters to be tuned
- No probabilistic interpretation
  - Hard to know how confident to be in each result

Possible Projects

- Tag based on answer types
- Do NER on question mentions “this English poet”