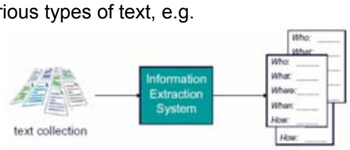
CSC 594 Topics in AI –
Applied Natural Language Processing

Fall 2009/2010

9. Information Extraction

Information Extraction (IE)

- Identify specific pieces of information (data) in an unstructured or semi-structured text
- Transform unstructured information in a corpus of texts or web pages into a structured database (or templates)
- Applied to various types of text, e.g.
  - Newspaper articles
  - Scientific articles
  - Web pages
  - etc.



Source: J. Choi, CSE842, MSU

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

template filling

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: **ACTIVITY-1**
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Why Information Extraction (IE)?

- Science
  - Grand old dream of AI: Build large knowledge base (KB) and reason with it. IE enables the automatic creation of this KB.
  - IE is a complex problem that inspires new advances in machine learning.
- Profit
  - Many companies interested in leveraging data currently “locked in unstructured text on the Web”.
  - Not yet a monopolistic winner in this space.
- Fun!
  - Build tools that we researchers like to use ourselves: Cora & CiteSeer, MRQE.com, FAQFinder,…
  - See our work get used by the general public.

Source: Andrew McCallum, UMass Amherst

A Typical IE Processing Pipeline

Named Entity Recognition (NER) &
Shallow Parsing

Reference Resolution

Relation Detection & Classification

Event Detection & Classification

Template Filling

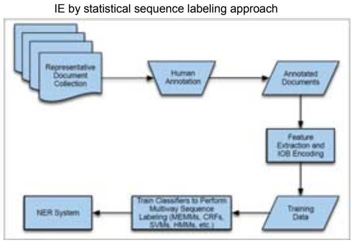
1. Named Entity Recognition

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

IOB notation

Word	POS	Chunk	EntityType
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

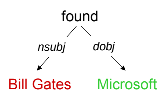
IE by statistical sequence labeling approach



1. Named Entity Recognition (cont.)

- After NEs are identified, **dependency parsing** is often applied to extract the syntactic/dependency relations between the NEs.

[$_{PER}$ Bill Gates] founded [$_{ORG}$ Microsoft].



Dependency Relations
nsubj(Bill Gates, found)
dobj(found, Microsoft)

2. Reference Resolution

- Two types of references:
  - Anaphora resolution
    - Identify what a **pronoun** refers to (an entity that appeared earlier in the text) – "he", "she", "it", "they"
  - Co-reference resolution
    - Identify what a noun (or noun phrase) refers to

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

- Reference resolution is an important step in IE and a very difficult problem in NLP. However, we don't cover it in this class.

3. Relation Detection

- Identify the semantic relations between named entities (or domain elements)
- Relations include:
  - General relations such as "part-of" and "employs"
  - Domain-specific relations

Semantic relations with examples and the NE types they involve

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geospatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

Source: Jurafsky & Martin "Speech and Language Processing"

3.1 Supervised Learning for Relation Analysis

- Training data:
  - Use a corpus annotated with NEs and relations
  - An instance indicates two arguments, their roles, and the type of the relation involved

Domain	$\mathcal{D} = \{a,b,c,d,e,f,g,h,i\}$
United, UAL, American Airlines, AMR	a,b,c,d
Tim Wagner	e
Chicago, Dallas, Denver, and San Francisco	f,g,h,i
Classes	
United, UAL, American, and AMR are organizations	$Org = \{a,b,c,d\}$
Tim Wagner is a person	$Pers = \{e\}$
Chicago, Dallas, Denver, and San Francisco are places	$Loc = \{f,g,h,i\}$
Relations	
United is a unit of UAL	$PartOf = \{\langle a,b\rangle, \langle c,d\rangle\}$
American is a unit of AMR	
Tim Wagner works for American Airlines	$OrgAff = \{\langle c,e\rangle\}$
United serves Chicago, Dallas, Denver, and San Francisco	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

Source: Jurafsky & Martin "Speech and Language Processing"

Supervised Learning for Relation Analysis (cont.)

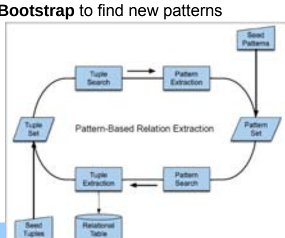
- Typical approach:
  - Step 1: Detect if a relation exists between two entities
  - Step 2: If so, classify/label the relation
- Features to represent an instance:

Entity-based features	
$Entity_1$ type	ORG
$Entity_1$ head	*airlines*
$Entity_2$ type	PERS
$Entity_2$ head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before $Entity_1$	NONE
Word(s) after $Entity_2$	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Source: Jurafsky & Martin "Speech and Language Processing"

3.2 Pattern-based Relation Analysis

- When an annotated corpus is not available, lightly-supervised methods can be used
  - Step 1: Define a set of **seed patterns** as **regular expressions**, and extract all tuples that match the patterns
  - Step 2: **Bootstrap** to find new patterns



4. Event Detection

- Identify events or states mentioned for named entities (or domain elements)

> [$_{event}$ Citing] high fuel prices, United Airlines [$_{event}$ said] Friday it has [$_{event}$ increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers.

- In English, events correspond to **verbs**, plus some noun phrases (e.g. "the increase", "the destruction")
  - But some difficult cases, e.g.
    - Not all verbs denote an event (e.g. "took effect")
    - 'Light verbs' such as "make", "take" are too generic. Must also look at its direct object/noun (e.g. "took a flight")
    - Of course, always the problem of (meaning) ambiguity…

Source: Jurafsky & Martin "Speech and Language Processing" 13

Event Detection (cont.)

- Both rule-based and statistical ML approaches have been used for event detection
- Features to represent an event instance:

Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character level suffixes for nominalizations (e.g., *-tion*)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
WordNet hypernyms	Hypernym set for the target

Source: Jurafsky & Martin "Speech and Language Processing" 14

5. Template Filling

- A template is a frame (of a record structure), consisting of slots and fillers. A template denotes an event or a semantic concept.
- After extracting NEs, relations and events, IE fills an appropriate template

FARE-RAISE ATTEMPT:		
	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

- Two common approaches for template filling:
  - Statistical approach
  - Finite-state cascade approach

Source: Jurafsky & Martin "Speech and Language Processing" 15

5.1 Statistical Approach to Template Filling

- Again, by using a sequence labeling method:
  - Label sequences of tokens as potential fillers for a particular slot
  - Train separate sequence classifiers for each slot
  - Slots are filled with the text segments identified by each slot's corresponding classifier
  - Resolve multiple labels assigned to the same/overlapping text segment by adding weights (heuristic confidence) to the slots
  - State-of-the-art performance – F1-measure of 75 to 98
- However, those methods are shown to be effective only for small, homogenous data.

Source: Jurafsky & Martin "Speech and Language Processing"

5.2 Finite-State Template-Filling Systems

- Message Understanding Conferences (MUC) – the genesis of IE
  - DARPA funded significant efforts in IE in the early to mid 1990's.
  - MUC was an annual event/competition where results were presented.
  - Focused on extracting information from news articles:
    - Terrorist events (MUC-4, 1992)
    - Industrial joint ventures (MUC-5, 1993)
    - Company management changes
  - Information extraction of particular interest to the intelligence community (CIA, NSA). (Note: early '90's)

Source: Marti Hearst, i256, at UC Berkeley

Finite-State Template-Filling Systems (cont.)

- FASTUS system in MUC-5
  - A cascade of transducers, where each level is a finite-state automata which extracts a specific type of information
  - The task was to fill hierarchically linked templates

No.	Step	Description
1	Tokens:	Transfer an input stream of characters into a token sequence.
2	Complex Words:	Recognize multiword phrases, numbers, and proper names.
3	Basic phrases:	Segment sentences into noun groups, verb groups, and particles.
4	Complex phrases:	Identify complex noun groups and complex verb groups.
5	Semantic Patterns:	Identify semantic entities and events and insert into templates.
6	Merging:	Merge references to the same entity or event from different parts of the text.

	Template/Slot	Value
1	RELATIONSHIP:	TIE-UP
	ENTITIES:	"Bridgestone Sports Co."
		"a local concern"
		"a Japanese trading house"
2	ACTIVITY:	PRODUCTION
	PRODUCT:	"golf clubs"
3	RELATIONSHIP:	TIE-UP
	JOINTVENTURECOMPANY:	"Bridgestone Sports Taiwan Co."
	AMOUNT:	NT$20000000
4	ACTIVITY:	PRODUCTION
	COMPANY:	"Bridgestone Sports Taiwan Co."
	STARTDATE:	DURING: January 1990
5	ACTIVITY:	PRODUCTION
	PRODUCT:	"iron and "metal wood" clubs"

Source: Marti Hearst, i256, at UC Berkeley

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: **ACTIVITY-1**
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Source: Marti Hearst, i256, at UC Berkeley

19

Evaluating IE Accuracy

- Precision and Recall:
  - Precision: correct answers / answers produced
  - Recall: correct answers / total possible correct answers
- F-measure:

$$F = \frac{(\beta^2 + 1)P * R}{(\beta^2 P + R)}$$

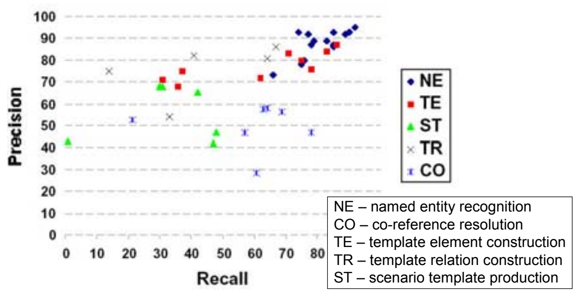
where β is a parameter representing relative importance of P and R.

When P and R are equally important, β = 1 and we get the F1 measure: $F1 = \frac{2 * P * R}{P + R}$

Source: J. Choi, CSE842, MSU

20

MUC Information Extraction: State of the Art c. 1997



Source: Marti Hearst, i256, at UC Berkeley

21

Successors to MUC

- CoNLL: Conference on Computational Natural Language Learning
  - Different topics each year
  - 2002, 2003: Language-independent NER
  - 2004: Semantic Role recognition
  - 2001: Identify clauses in text
  - 2000: Chunking boundaries
    - http://cnts.uia.ac.be/conll2003/ (also conll2004, conll2002…)
    - Sponsored by SIGNLL, the Special Interest Group on Natural Language Learning of the Association for Computational Linguistics.
- ACE: *Automated Content Extraction*
  - Entity Detection and Tracking
    - Sponsored by NIST
    - http://wave.ldc.upenn.edu/Projects/ACE/
- Several others recently
  - See http://cnts.uia.ac.be/conll2003/ner/

Source: Marti Hearst, i256, at UC Berkeley

22

State of the Art Performance: examples

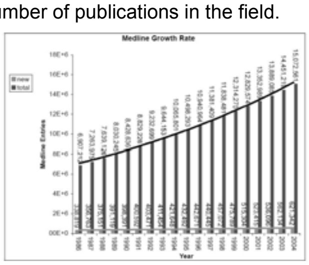
- Named entity recognition from newswire text
  - Person, Location, Organization, …
  - F1 in high 80's or low- to mid-90's
- Binary relation extraction
  - Contained-in (Location1, Location2)
    Member-of (Person1, Organization1)
  - F1 in 60's or 70's or 80's
- Web site structure recognition
  - Extremely accurate performance obtainable
  - Human effort (~10min?) required on each site

Source: Marti Hearst, i256, at UC Berkeley

23

Advanced Topic: Biomedical Information Extraction

- IE from biomedical journal articles has become an important application area lately – due to a rapid growth in the number of publications in the field.



Source: Jurafsky & Martin "Speech and Language Processing"

24

Biological NER

- There are a much wider range of entity types (semantic classes) in the biological domain

> [$_{\text{tissue}}$ Plasma] [$_{\text{GP}}$ BNP] concentrations were higher in both the [$_{\text{population}}$ judo] and [$_{\text{population}}$ marathon groups] than in [$_{\text{population}}$ controls], and positively correlated with [$_{\text{ANAT}}$ LV] mass as well as with deceleration time.

Semantic class	Examples
Cell lines	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*
Cell types	*primary T lymphocytes, natural killer cells, NK cells*
Chemicals	*citric acid, 1,2-diiodopentane, C*
Drugs	*cyclosporin A, CDDP*
Genes/proteins	*white, HSP60, protein kinase C, L23A*
Malignancies	*carcinoma, breast neoplasms*
Medical/clinical concepts	*amyotrophic lateral sclerosis*
Mouse strains	*LAFT, AKR*
Mutations	*C10T, Ala64 → Gly*
Populations	*judo group*

Source: Jurafsky & Martin “Speech and Language Processing” 25

Biological NER (cont.)

- NER in this domain is particularly difficult because of the various forms which the names can take:
  e.g. “insulin”, “ether a go-go”, “breast cancer associated 1”
  - Long names (thus multi-token boundary detection is needed)
  - Spelling/typographical variations
  - Abbreviations, symbols
  - (Of course) Ambiguity (common meaning or domain concepts)
- Extracted NEs are often mapped to **biomedical ontologies** (e.g. Gene Ontology, UMLS)

Source: Jurafsky & Martin “Speech and Language Processing” 26

Biological Roles and Relations

- Two approaches:
  - Discover and classify binary relations between NEs

> (1) These results suggest that con A-induced [$_{\text{disease}}$ hepatitis] was ameliorated by pretreatment with [$_{\text{treatment}}$ TJ-135]. ➔ ‘curing’ relation
>
> (2) [$_{\text{disease}}$ Malignant mesodermal mixed tumor of the uterus] following [$_{\text{treatment}}$ irradiation] ➔ ‘result’ relation

  - Identify and classify the roles played by NEs w.r.t. the event --- which constituents for which semantic roles

> [$_{\text{theme}}$ Full-length cPLA2] was [$_{\text{target}}$ phosphorylated] stoichiometrically by [$_{\text{agent}}$ p43 mitrogen-activated protein (MAP) kinase] into vitro … and the major site of phosphorylation was identified by amino acid sequencing as [$_{\text{site}}$ Ser505].

Note: the event for this sentence is ‘PHOSPHORYLATION’

Source: Jurafsky & Martin “Speech and Language Processing” 27

Automatic Role Labeling for Biological Domain

- Both rule-based and statistical approaches have been applied
- Medical ontologies (in particular the link/inference structures) are often utilized
- General results: **The choice of algorithm is less important than the choice of features**
- Note: NER methods utilize syntactic features -- but no large treebanks are available for biomedical domain ➔ Off-the-shelf NER tools (trained with generic newswire exts) are often used.

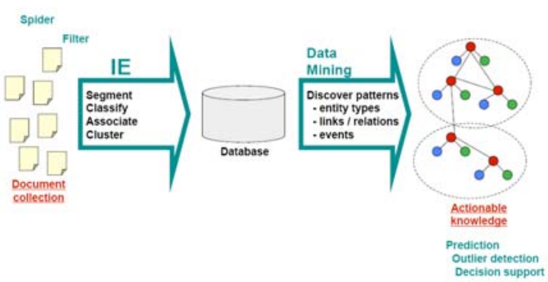
Source: Jurafsky & Martin "Speech and Language Processing" 28

IE Techniques: Summary

- Machine learning approaches are doing well, even without comprehensive word lists
  - Can develop a pretty good starting list with a bit of web page scraping
  - Lately **Conditional Random Fields (CRFs)** have shown superb performance over other sequence-labeling ML techniques
- Features mainly have to do with the preceding and following tags, as well as syntax and orthographic features of words
  - The latter is somewhat language dependent
- With enough training data, results are getting pretty decent on well-defined entities
- ML is the way of the future!

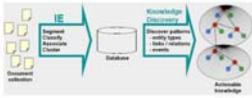
Source: Marti Hearst, i256, at UC Berkeley 29

Extra: From Text to Actionable Knowledge



Source: Andrew McCallum, UMass Amherst 30

Problem:

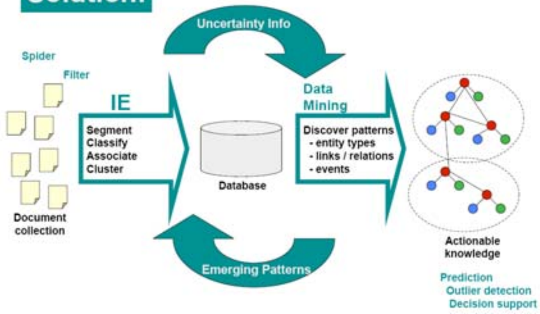


Combined in serial juxtaposition, IE and DM are unaware of each others' weaknesses and opportunities.

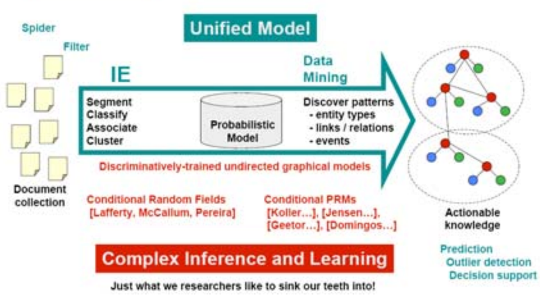
1) DM begins from a populated DB, unaware of where the data came from, or its inherent errors and uncertainties.
2) IE is unaware of emerging patterns and regularities in the DB.

The accuracy of both suffers, and significant mining of complex text sources is beyond reach.

Solution:



Solution:



Research Questions

- What model structures will capture salient dependencies?
- Will joint inference actually improve accuracy?
- How to do ***inference*** in these large graphical models?
- How to do ***parameter estimation*** efficiently in these models, which are built from multiple large components?
- How to do ***structure discovery*** in these models?

Source: Andrew McCallum, UMass Amherst

34