COM6115:
Text Processing

Introduction to Information Extraction

Chenghua Lin

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: Rule-based; Supervised Learning
  - Entity Linking
- Relation Extraction
  - Task
  - Approaches: Rule-based; Supervised learning; Bootstrapping; Distant Supervision

Introduction to Information Extraction: Outline

- Definition + Contrast with IR
- Example Applications
- Overview of Tasks
- Overview of Approaches
- Evaluation + Shared Task Challenges
- A Brief History of IE

Definition

- Definition: the Information Extraction (IE) task:

  From each text in a set of unstructured natural language texts identify information about predefined classes of entities, relationships or events and record this information in a structured form by either:

  - Annotating the source text, e.g. using XML tags; or
  - Filling in a data structure separate from the text, e.g a template or database record or "stand-off annotation"

- For example: from financial newswire stories identify those dealing with management succession events and from these extract details of organisations and persons, the post being assumed or vacated, etc.

Definition (cont)

- IE may also be described as:
  - The activity of populating a structured information repository (database) from an unstructured, or free text, information source
  - The activity of creating a semantically annotated text collection (cf. "The Semantic Web")
- The resulting structured data source is then used for some other purpose:
  - searching or analysis using conventional database queries;
  - data-mining;
  - generating a summary (perhaps in another language);

Example

Who's News: @ Burns Fry Ltd.
04/13/94
WALL STREET JOURNAL (J), PAGE B10 ⟨CO⟩
BURNS FRY Ltd. (Toronto) – Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Example

Who's News: @ Burns Fry Ltd.
04/13/94
WALL STREET JOURNAL (J), PAGE B10 $\langle$CO$\rangle$
BURNS FRY Ltd. (Toronto) – Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

- identify persons (red)

Example

Who's News: @ Burns Fry Ltd.
04/13/94
WALL STREET JOURNAL (J), PAGE B10 ⟨CO⟩
BURNS FRY Ltd. (Toronto) – Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

- identify persons (red)
- identify organisations (blue)

Example

Who's News: @ Burns Fry Ltd.
04/13/94
WALL STREET JOURNAL (J), PAGE B10 ⟨CO⟩
BURNS FRY Ltd. (Toronto) – Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

- identify persons (red)
- identify organisations (blue)
- identify locations (green)

Example

Who's News: @ Burns Fry Ltd.
04/13/94
WALL STREET JOURNAL (J), PAGE B10 ⟨CO⟩
BURNS FRY Ltd. (Toronto) – Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

- identify persons (red)
- identify organisations (blue)
- identify locations (green)
- identify times (cyan)

Example

Who's News: @ Burns Fry Ltd.
04/13/94
WALL STREET JOURNAL (J), PAGE B10 ⟨CO⟩
BURNS FRY Ltd. (Toronto) – Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

- identify persons (red)
- identify organisations (blue)
- identify locations (green)
- identify times (cyan)
- identify company positions (purple)

Example

Who's News: @ Burns Fry Ltd.
04/13/94
WALL STREET JOURNAL (J), PAGE B10 ⟨CO⟩
BURNS FRY Ltd. (Toronto) – Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

- identify persons (red)
- identify organisations (blue)
- identify locations (green)
- identify times (cyan)
- identify company positions (purple)
- identify succession events (underlined)

Example: Filled Template

```
<TEMPLATE-9404130062> :=
   DOC_NR: "9404130062"
   CONTENT: <SUCCESSION_EVENT-1>
<SUCCESSION_EVENT-1> :=
   SUCCESSION_ORG: <ORGANIZATION-1>
   POST: "executive vice president"
   IN_AND_OUT:   <IN_AND_OUT-1> <IN_AND_OUT-2>
   VACANCY_REASON: OTH_UNK
<IN_AND_OUT-1> :=
   IO_PERSON: <PERSON-1>
   NEW_STATUS: OUT
   ON_THE_JOB: NO
<ORGANIZATION-1> :=
   ORG_NAME: "Burns Fry Ltd.“
   ORG_ALIAS: "Burns Fry“
   ORG_DESCRIPTOR: "this brokerage firm“
   ORG_TYPE: COMPANY
   ORG_LOCALE: Toronto CITY
   ORG_COUNTRY: Canada
<PERSON-1> :=
   PER_NAME: "Mark Kassirer"

<IN_AND_OUT-2> :=
   IO_PERSON: <PERSON-2>
   NEW_STATUS: IN
   ON_THE_JOB: NO
   OTHER_ORG: <ORGANIZATION-2>
   REL_OTHER_ORG: OUTSIDE_ORG
<ORGANIZATION-2> :=
   ORG_NAME: "Merrill Lynch Canada Inc."
   ORG_ALIAS:  "Merrill Lynch"
   ORG_DESCRIPTOR: "a unit of Merrill Lynch & Co."
   ORG_TYPE: COMPANY
<PERSON-2> :=
   PER_NAME: "Donald Wright"
```

Contrast with Information Retrieval

Information Retrieval

- Task:
  - ◇ Given: a document collection and a user query
  - ◇ Return: a (ranked) list of documents relevant to the user query
- Strengths:
  - ◇ Can search huge document collections very rapidly
  - ◇ Insensitive to genre and domain of the texts
  - ◇ Relatively straightforward to implement
    - challenges scaling to huge, dynamic document collections, e.g. the Web
- Weaknesses
  - ◇ Documents are returned not information/answers, so
    - user must further read texts to extract information
    - output is unstructured so limited possibilities for direct data mining/further processing

Contrast with Information Retrieval

Information Extraction

- Task:
  - Given: a document collection and a predefined set of entities, relations and/or events
  - Return: a structured representation of all mentions of the specified entities, relations and/or events
- Strengths:
  - Extracts facts from texts, not just texts from text collections
  - Can feed other powerful applications (databases, semantic indexing engines, data mining tools)
- Weaknesses
  - Systems tend to be genre/domain specific and porting to new genres and domains can be time-consuming/requires expertise
  - Limited accuracy
  - Computationally demanding, so performance issues on very large collections

Example Applications

- Scrapping web pages to build structured databases of job postings, apartment rentals, seminar announcements, etc.
- Assisting biomedical database curators by extracting biomedical entities and relations from the scientific literature prior to entry in a human-maintained database (e.g. Flybase)
- Assisting companies in competitor intelligence gathering, e.g. management or researcher succession events, new product or project annoucements, etc.

Introduction to Information Extraction: Outline

- Definition + Contrast with IR
- Example Applications
- Overview of Tasks
- Overview of Approaches
- Evaluation + Shared Task Challenges
- A Brief History of IE

Overview of Tasks: Entity Extraction

Entity Extraction/Named Entity Recognition

- **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 ...

Overview of Tasks: Entity Extraction – Coreference

- 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

Relation Extraction

- **Task**: identify all assertions of relations, usually binary, between entities identified in entity extraction
- May be divided into two subtasks:
  - **Relation detection**: find pairs of entities between which a relation holds
  - **Relation classification**: for pairs of entities between which a relation holds, determine what the relation is
- Examples
  - LOCATION_OF holding between
    - ORGANISATION and GEOPOLITICAL_LOCATION
    - medical INVESTIGATION and BODY_PART
  - EMPLOYEE_OF holding between PERSON and ORGANISATION
  - PRODUCT_OF holding between ARTIFACT and ORGANISATION
  - INTERACTION holding between PROTEIN and PROTEIN

Overview of Tasks: Relation Extraction

Relation Extraction is challenging for several reasons:

- The same relation may be expressed in many different ways:
  - ◇ Synonyms: [Microsoft]$_{ORG}$ is based/headquartered in [Redmond]$_{LOC}$
  - ◇ Syntactic variations:
    - [Microsoft]$_{ORG}$, the software giant and ..., is based in [Redmond]$_{LOC}$
    - [Redmond]$_{LOC}$-based [Microsoft]$_{ORG}$ ...
    - [Redmond]$_{LOC}$'s [Microsoft]$_{ORG}$ ...; [Microsoft]$_{ORG}$ of [Redmond]$_{LOC}$
    - [Redmond]$_{LOC}$ software giant [Microsoft]$_{ORG}$ ...
- Discovering relations frequently depends upon being able to follow coreference links.
  Dirk Ruthless of MegaCorp made a stunning announcement today. In September he will be stepping down as Chief Executive Officer to spend more time with his pet piranhas.
  To determine the corporate position of Dirk Ruthless we must correctly resolve the pronominal anaphor "he" in the second sentence with "Dirk Ruthless" in the first

Overview of Tasks: Event Detection

Event Extraction

- **Task**: identify all reports of event instances, typically of a small set of classes
- May be divided into two subtasks:
  - ◇ **Event detection**: find mentions of events in text
  - ◇ **Event classification**: assign detected events to one of a set of classes
- Examples
  - ◇ Rocket/missile launches
  - ◇ Management succession events
  - ◇ Joint venture/product announcements
  - ◇ Terrorist attacks
- Events may be simply viewed as relations. However they are typically complex relations that
  - ◇ Are temporally situated and often of relatively short duration
  - ◇ Involve multiple role players (frequently > 2)
  - ◇ Are often expressed across multiple sentences

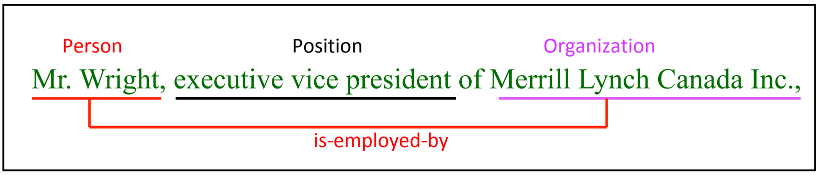
Introduction to Information Extraction: Outline

- Definition + Contrast with IR
- Example Applications
- Overview of Tasks
- Overview of Approaches
- Evaluation + Shared Task Challenges
- A Brief History of IE

Overview of Approaches:

Approaches to IE may be placed into four categories:

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

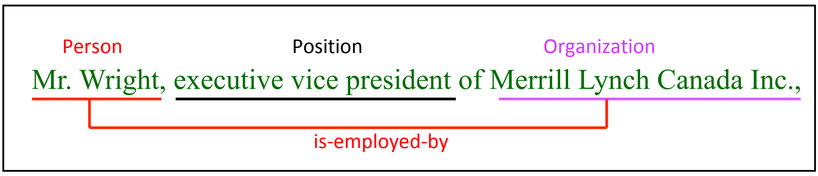


- Such systems use manually authored rules and can be divided into
  - “deep” – linguistically inspired “language understanding” systems
  - “shallow” – systems engineered to the IE task, typically using pattern-action rules

    Pattern: `‘‘Mr. $Uppercase-initial-word’’`
    Action: add-entity(person(“Mr. $Uppercase-initial-word))

    Pattern: `\$Person, $Position of $Organization"`
    Action: add-relation(is-employed-by($Person,$Organization))

Supervised learning approaches



- Systems are given texts with manually annotated entities + relations
- For each entity/relation create a training instance
  - k words either side of an entity mention
  - k words to the left of entity 1 and to the right of entity 2 plus the words in between
- Training instances represented in terms of features
  - words, parts of speech, orthographic characteristics, syntactic info
- Systems may learn
  - patterns that match extraction targets
  - Classifiers that classify tokens as beginning/inside/outside a tag type
- Learning techniques include: covering algorithms, HMMs, SVMs

Bootstrapping Approaches

- A technique for relation extraction that requires only minimal supervision
- Systems are given
  - seed tuples (e.g. ⟨ Microsoft, Redmond ⟩)
  - seed patterns (e.g. $[X]_{ORG}$ is located in $[Y]_{LOC}$)

  or both.
- System searches in large corpus for
  - occurences of seed tuples and then extracts a pattern that matches the context of the seed tuple
  - matches of seed patterns from which it harvests new tuples
- New tuples are assumed to stand in the required relation and are added to the tuple store
- Process iterates until convergence
- See later lecture

Distant Supervision Approaches

- Sometimes also called “weakly labelled” approaches
- Assumes a (semi-)structured data source, such as
  - Wikipedia infoboxes (e.g. PERSON BORN_IN LOCATION/DATE)
  - Freebase or Wikidata
  - Flybase or the Yeast Protein Database, (e.g. PROTEIN IS_LOCATED_IN SUBCELLULAR_LOCATION)

  which contains tuples of entities standing in the relation of interest and, ideally, a pointer to a source text
- Tuples from data source are used to label
  - the text with which they are associated, if available
  - documents from the web, if not
- Labelled data is used to train a standard supervised named entity or relation extraction system
- See later lecture

Introduction to Information Extraction: Outline

- Definition + Contrast with IR
- Example Applications
- Overview of Tasks
- Overview of Approaches
- Evaluation + Shared Task Challenges
- A Brief History of IE

Evaluation

- Correct answers, called keys, are produced manually for each extraction task (filled templates or SGML annotated texts)
- Scoring of system results, called responses, against keys is done automatically.
- At least some portion of the answer keys are multiply produced by different humans so that interannotator agreement figures can be computed.
- Principal metrics – borrowed from information retrieval – are:
  - ◇ Precision (how much of what system returns is correct)
  - ◇ Recall (how much of what is correct system returns)
  - ◇ F-measure (a weighted combination of precision and recall)

Shared Task Challenges

- Shared Task challenges are community wide exercises in which groups of researchers engage in a friendly competition to build systems to address a common task
- Key elements are:
  - an agreed task definition
  - annotated text resources for training and testing
  - agreed metrics for evaluation
  - an agreed schedule for release of resources, system development, system evaluation and a conference to discuss results
- Shared task challenges in IE have included: MUC, ACE, TAC, BioCreative
- Define the core methodology of the field and have led to significant progress

Introduction to Information Extraction: Outline

- Definition + Contrast with IR
- Example Applications
- Overview of Tasks
- Overview of Approaches
- Evaluation + Shared Task Challenges
- A Brief History of IE

A Brief History of IE

1960s The first published work on information extraction (though it was not called this at the time)

1970s A significant precursor was the psychologist Roger Schank's work on scripts and story understanding

1980s Saw the emergence of some commercial systems targetted at financial transactions and newswires

Message Understanding Conference 1 (MUC-1) – in 1987

1990s MUC ran 7 times until 1998 and significantly advanced the field.

Machine learning approaches to IE began to appear

2000s ACE (Automatic Content Extraction) the successor programme to MUC ran 1999-2008; succeeded by TAC (Text Analytics Conference) (2008-present); BioCreative (IE in the biomedical domain) began (2004-present); work on IE in other languages began (e.g. Spanish, Japanese, Chinese, Arabic)

2010s TAC is going, particularly the knowledge base population track

Currently there are a number of IE systems on the market and a large and on-going research effort in the field