Processing of large document collections

Part 8 (Information extraction)
Helena Ahonen-Myka
Spring 2005

5. Information extraction

- in this part:
  - task definition
  - information extraction (IE) compared to other related fields
  - generic IE process

Reference

- the following is largely based on
  - Ralph Grishman: Information extraction: Techniques and Challenges. In Information Extraction, a multidisciplinary approach to an emerging information technology. Lecture Notes in AI, Springer, 1997.

Task

- ”Information extraction involves the creation of a structured representation (such as a database) of selected information drawn from the text”

Example: terrorist events

19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Example: terrorist events

- a document collection is given
- for each document, decide if the document is about terrorist event
- for each terrorist event, determine
  - type of attack
  - date
  - location, etc.
- = fill in a template (~database record)

Example: terrorist events

Incident type	bombing
Date	March 19
Location	El Salvador: San Salvador (city)
Perpetrator	urban guerilla commandos
Physical target	power tower
Human target	-
Effect on physical target	destroyed
Effect on human target	no injury or death
Instrument	bomb

7

Message understanding conferences (MUC)

- development of IE systems has been shaped by a series of evaluations, the MUC conferences (1987-98)
- MUCs have provided IE tasks and sets of training and test data + evaluation procedures and measures
- participating projects have competed with each other but also shared ideas

8

Other tasks (in MUC)

- international joint ventures
  - facts to be found: partners, the new venture, its product or service, etc.
- executive succession
  - who was hired/fired by which company for which position

9

IE compared to other related fields

- IE vs. information retrieval
- IE vs. full text understanding

10

IE vs. information retrieval

- Information retrieval (IR)
  - given a user query, an IR system selects a (hopefully) relevant subset of documents from a larger set
  - the user then browses the selected documents in order to fulfil his or her information need
- IE extracts relevant information from documents -> IR and IE are complementary technologies

11

IE vs full text understanding

- in text understanding
  - the aim is to make sense of the entire text
  - the target representation must accommodate the full complexities of language
  - one wants to recognize the nuances of meaning and the writer's goals

12

IE vs full text understanding

- in IE
  - generally only a fraction of the text is relevant
  - information is mapped into a predefined, relatively simple, rigid target representation
  - the subtle nuances of meaning and the writer's goals in writing the text are of secondary interest

13

Generic IE process

- rough view of the IE process:
  - the system extracts individual "facts" from the text of a document through local text analysis
  - the system integrates these facts, producing larger facts or new facts (through inference)
  - the facts are translated into the required output format

14

Process: more detailed view

- the individual facts are extracted by creating a set of patterns to match the possible linguistic realizations of the facts
  - it is not practical to describe these patterns directly as word sequences
  - the input is structured; various levels of constituents and relations are identified
  - the patterns are stated in terms of these constituents and relations

15

Process: stages

- local text analysis phase (separately for each sentence):
  - 1. lexical analysis
    - assigning part-of-speech and other features to words/phrases through morphological analysis and dictionary lookup
  - 2. name recognition
    - identifying names and other special lexical structures such as dates, currency expressions, etc.

16

Process: stages

  - 3. full syntactic analysis or some form of partial parsing
    - partial parsing: e.g. identify noun groups, verb groups
  - 4. task-specific patterns are used to identify the facts of interest

17

Process: stages

- integration phase: examines and combines facts from the entire document
  - 5. coreference analysis
    - use of pronouns, multiple descriptions of the same event
  - 6. inferencing from the explicitly stated facts in the document

18

Some terminology

- domain
  - general topical area (e.g. financial news)
- scenario
  - specification of the particular events or relations to be extracted (e.g. joint ventures)
- template
  - final, tabular (record) output format of IE
- template slot, argument (of a template)
  - e.g. location, human target

19

Running example

- ”Sam Schwartz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. He will be succeeded by Harry Himmelfarb.”

20

Target templates

Event	leave job
Person	Sam Schwartz
Position	executive vice president
Company	Hupplewhite Inc.

Event	start job
Person	Harry Himmelfarb
Position	executive vice president
Company	Hupplewhite Inc

21

Lexical analysis

- the text is divided into sentences and into tokens (”words”)
- each token is looked up in the dictionary to determine its possible part of speech and features
  - general-purpose dictionaries
  - special dictionaries
    - major place names, major companies, common first names, company suffixes (”Inc.”)

22

Lexical analysis

- Sam: known first name -> person
- Schwartz: unknown capitalized word
- retired: verb
- as: preposition
- executive: adjective
- vice: adjective
- president: noun (person?)

23

Name recognition

- various types of proper names and other special forms, such as dates and currency amounts, are identified and classified
  - classes e.g. person name, company name
- names appear frequently in many types of texts: identifying and classifying them simplifies further processing
  - instead of several distinct words, the whole name can be processed as one entity
- names are also important as template slot values for many extraction tasks

24

Name recognition

- names are identified by a set of patterns (regular expressions) which are stated in terms of part of speech, syntactic features, and orthographic features (e.g. capitalization)

25

Name recognition

- personal names might be identified
  - by a preceding title: Mr. Herrington Smith
  - by a common first name: Fred Smith
  - by a suffix: Snippety Smith Jr.
  - by a middle initial: Humble T. Hopp

26

Name recognition

- company names can usually be identified by their final token(s), such as
  - Hepplewhite Inc.
  - Hepplewhite Corporation
  - Hepplewhite Associates
  - First Hepplewhite Bank
- however, some major company names (”General Motors”) are problematic
  - dictionary of major companies is needed

27

Name recognition

- <name type=”person”> Sam Schwartz </name> retired as executive vice president of the famous hot dog manufacturer, <name type=”company”> Hupplewhite Inc.</name>
- He will be succeeded by <name type=”person”>Harry Himmelfarb</name>.

28

Name recognition

- subproblem: identify the aliases of a name (name coreference)
  - Larry Liggett = Mr. Liggett
  - Hewlett-Packard Corp. = HP
- alias identification may also help name classification
  - ”Humble Hopp reported…” (person or company?)
  - subsequent reference: ”Mr. Hopp” (-> person)

29

Syntactic analysis

- identifying syntactic structure:
  - “grouping words” , forming phrases
    - noun phrases: sam schwartz, executive vice president; approximately 5 kg, more than 30 peasants
    - verb groups: retired, will be succeeded
  - finding grammatical functional relations
    - subject, (direct/indirect) object, main verb

30

Syntactic analysis

- identifying some aspects of syntactic structure simplifies the subsequent phase of fact extraction
  - the slot values to be extracted often correspond to noun phrases
  - the relationships often correspond to grammatical functional relations
- but: identification of the complete syntactic structure of a sentence is difficult

31

Syntactic analysis

- problems e.g. with prepositional phrases to the right of a noun
  - ”I saw the man in the park with a telescope.”
    - the prepositional phrases can be associated both with ”man” and with ”saw”

32

Syntactic analysis

- in extraction systems, there is a great variation in the amount of syntactic structure which is explicitly identified
  - some systems do not have any separate phase of syntactic analysis
  - others attempt to build a complete parse of a sentence
  - most systems fall in between and build a series of parse fragments

33

Syntactic analysis

- systems that do partial parsing
  - build structures about which they can be quite certain, either from syntactic or semantic evidence
    - for instance, structures for noun groups (a noun + its left modifiers) and for verb groups (a verb with its auxiliaries)
    - both can be built using just local syntactic information
    - in addition, larger structures can be built if there is enough semantic information

34

Syntactic analysis

- the first set of patterns labels all the basic noun groups as noun phrases (np)
- the second set of patterns labels the verb groups (vg)

35

Syntactic analysis

- <np entity=”e1”> Sam Schwartz </np> <vg>retired</vg> as <np entity=”e2”> executive vice president</np> of <np entity=”e3”>the famous hot dog manufacturer</np>, <np entity=”e4”> Hupplewhite Inc.</np>
- <np entity=”e5”>He</np> <vg>will be succeeded</vg> by <np entity=”e6”>Harry Himmelfarb</np>.

36

Syntactic analysis

- associated with each constituent are certain features which can be tested by patterns in subsequent stages
  - for verb groups: tense (past/present/future), voice (active/passive), baseform/stem
  - for noun phrases: baseform/stem, is this phrase a name?, number (singular/plural)

37

Syntactic analysis

- For each NP, the system creates a semantic entity

entity e1	type: person	name: ”Sam Schwartz”
entity e2	type: position	value: ”executive vice president”
entity e3	type: manufacturer	
entity e4	type: company	name:”Hupplewhite Inc.”
entity e5	type: person	
entity e6	type: person	name: ”Harry Himmelfarb”

38

Syntactic analysis

- semantic constraints
  - the next set of patterns build up larger noun phrase structures by attaching right modifiers
  - because of the syntactic ambiguity of right modifiers, these patterns incorporate some semantic constraints (domain specific)

39

Syntactic analysis

- in our example, two patterns will recognize the appositive construction:
  - *company-description, company-name*,
- and the prepositional phrase construction:
  - *position* of *company*
- in the second pattern:
  - *position* matches any NP whose entity is of type ”position”
  - *company* respectively

40

Syntactic analysis

- the system includes a small semantic type hierarchy (*is-a* hierarchy)
  - e.g. manufacturer is-a company
  - the pattern matching uses the *is-a* relation, so any subtype of company (such as manufacturer) will be matched

41

Syntactic analysis

- in the first pattern
  - *company-name*: NP of type ”company” whose head is a name
    - e.g. ”Hupplewhite Inc.”
  - *company-description*: NP of type ”company” whose head is a common noun
    - e.g. ”the famous hot dog manufacturer”

42

Syntactic analysis

- after the first pattern is matched:
  - 2 NPs combined into one: the famous hot dog manufacturer, Hupplewhite Inc.
- further, after the second pattern:
  - executive vice president of the famous hot dog manufacturer, Hupplewhite Inc.
  - a new NP + the relationship between the position and the company

43

Syntactic analysis

- <np entity="e1"> Sam Schwartz </np> <vg>retired</vg> as <np entity="e2"> executive vice president of the famous hot dog manufacturer, Hupplewhite Inc.</np>
- <np entity="e5">He</np> <vg>will be succeeded</vg> by <np entity="e6"> Harry Himmelfarb</np>.

44

Syntactic analysis

- Entities are updated as follows:

entity e1	type: person	name: "Sam Schwartz"
entity e2	type: position	value: "executive vice president"
		company: e3
entity e3	type: manufacturer	name: "Hupplewhite Inc."
entity e5	type: person	
entity e6	type: person	name: "Harry Himmelfarb"

45

Scenario pattern matching

- role of scenario patterns is to extract the events or relationships relevant to the scenario
- in our example, there will be 2 patterns
  - *person* retires as *position*
  - *person* is succeeded by *person*
- *person* and *position* are pattern elements which match NPs with the associated type
- "retires" and "is succeeded" are pattern elements which match active and passive verb groups, respectively

46

Scenario pattern matching

- *person* retires as *position*
  - Sam Schwartz **retired as** executive vice president of the famous hot dog manufacturer, Hupplewhite Inc.
  - -> event leave-job (person, position)
- *person* is succeeded by *person*
  - He **will be succeeded by** Harry Himmelfarb
  - -> event succeed (person, person)

47

Scenario pattern matching

entity e1	type: person	name: "Sam Schwartz"
entity e2	type: position	value: "executive vice president"
		company: e3
entity e3	type: manufacturer	name:"Hupplewhite Inc."
entity e5	type: person	
entity e6	type: person	name: "Harry Himmelfarb"
event e7	type: leave-job	person: e1 position: e2
event e8	type: succeed	person1: e6 person2: e5

48

Scenario patterns for terrorist attacks

- for instance, in Fastus IE system, 95 scenario patterns
  - killing of <HumanTarget>
  - <GovOfficial> accused <PerpOrg>
  - bomb was placed by <Perp> on <PhysicalTarget>
  - <Perp> attacked <HumanTarget>'s <PhysicalTarget> with <Device>
  - <HumanTarget> was injured

49

Coreference analysis

- task of resolving anaphoric references by pronouns and definite noun phrases
  - in our example: "he" (entity e5)
  - coreference analysis will look for the most recent previously mentioned entity of type person, and will find entity e1
  - references to e5 are changed to refer to e1 instead
- also the *is-a* hierarchy is used

50

Coreference analysis

entity e1	type: person	name: "Sam Schwartz"
entity e2	type: position	value: "executive vice president"
		company: e3
entity e3	type: manufacturer	name:"Hupplewhite Inc."
entity e6	type: person	name: "Harry Himmelfarb"
event e7	type: leave-job	person: e1 position: e2
event e8	type: succeed	person1: e6 person2: e1

51

Inferencing and event merging

- partial information about an event may be spread over several sentences
  - this information needs to be combined before a template can be generated
- some of the information may also be implicit
  - this information needs to be made explicit through an inference process

52

Target templates?

Event	leave job
Person	Sam Schwartz
Position	executive vice president
Company	Hupplewhite Inc.
Event	
Person	Harry Himmelfarb
Position	
Company	

53

Inferencing and event merging

- in our example, we need to determine what the "succeed" predicate implies, e.g.
- "Sam was president. He was succeeded by Harry."
  - -> Harry will become president
- "Sam will be president; he succeeds Harry"
  - -> Harry was president.

54

Inferencing and event merging

- such inferences can be implemented by production rules:
  - leave-job(X-person,Y-job) & succeed(Z-person,X-person) => start-job(Z-person,Y-job)
  - start-job(X-person,Y-job) & succeed(X-person,Z-person) => leave-job(Z-person,Y-job)

55

Inferencing and event merging

entity e1	type: person	name: ”Sam Schwartz”
entity e2	type: position	value: ”executive vice president”
		company: e3
entity e3	type: manufacturer	name:”Hupplewhite Inc.”
entity e6	type: person	name: ”Harry Himmelfarb”
event e7	type: leave-job	person: e1 position: e2
event e8	type: succeed	person1: e6 person2: e1
event e9	type: start-job	person: e6 position:e2

56

Target templates

Event	leave job
Person	Sam Schwartz
Position	executive vice president
Company	Hupplewhite Inc.
Event	start job
Person	Harry Himmelfarb
Position	executive vice president
Company	Hupplewhite Inc.

57

Inferencing and event merging

- our simple scenario did not require us to take account of the time of each event
- for many scenarios, time is important
  - explicit times must be reported, or
  - the sequence of events is significant
- time information may be derived from many sources

58

Inferencing and event merging

- sources of time information
  - absolute dates and times (”on April 6, 1995”)
  - relative dates and times (”last week”)
  - verb tenses
  - knowledge about inherent sequence of events
- since time analysis may interact with other inferences, it will normally be performed as part of the inference stage of processing

59

Summary of IE process

- local analysis (for each sentence)
  - lexical analysis
  - name recognition
  - (partial) syntactic analysis
  - scenario pattern matching
- integration phase
  - coreference analysis
  - inferencing and event merging

60