Information Retrieval

Lecture 5: Information Extraction

Computer Science Tripos Part II

UNIVERSITY OF CAMBRIDGE

Stephen Clark

Natural Language and Information Processing (NLIP) Group

sc609@cl.cam.ac.uk

Information Extraction: the task

- Identify instances of a particular class of events or relationships in a natural language text
- Limited semantic range of events/relationships (domain-dependence)
- Extract the relevant arguments of the event or relationship into pre-existing “templates” (tabular data structures)
- MUC (Message Understanding Conference; NIST) 1986-97 competitive evaluation

History of IE

- 1980s and before: lexico-semantic patterns written by hand (FRUMP, satellite reports, patient discharge summaries...)
- 1987 First MUC (Message Understanding Conference); domain: naval sightings
- 1889 Second MUC; domain: naval sightings
- 1991 Third MUC; domain: terrorist acts
  - Winner (SRI) used partial parsing
- 1992 Fourth MUC; domain: terrorist acts
- 1993 Fifth MUC; domain: joint ventures/electronic circuit fabrication
  - Performance of best systems $\sim$ 40% R, 50% P (Humans in 60-80% range)
  - Lehnert et al.: first bootstrapping method

History of IE, ctd.

- 1995 Sixth MUC; domain: labour unit contract negotiations/changes in corporate executive management personnel
  - Encourage more portability and deeper understanding
  - Separate tasks into
    - NE: Named Entity
    - CO: Coreference
    - TE: Template Element
    - ST: Scenario Templates
- 1995: IE for summarisation (Radev and McKeown)
- 1998: Seventh MUC; domain: satellite rocket launch events
  - Mikheev et al., hybrid methods for NE
- 2003: CoNLL NE recognition task; similar training data to MUC

MUC setup

- Participants get a description of the scenario and a training corpus (a set of documents and the templates to be extracted from these)
- 1-6 months time to adapt systems to the new scenario
- NIST analysts manually fill templates of test corpus (“answer key”)
- Test corpus delivered; systems run at home
- Automatic comparison of system response with answer key
- Primary scores: precision and recall
- Participants present paper at conference in spring after competition
- Show system’s workings on predefined “walk through” example

Template example (MUC-3)

0	MESSAGE ID	TST1-MUC3-0080
1	TEMPLATE ID	1
2	DATE OF INCIDENT	03 APR 90
3	TYPE OF INCIDENT	KIDNAPPING
4	CATEGORY OF INCIDENT	TERRORIST ACT
5	PERPETRATOR: ID OF INDIV(S)	“THREE HEAVILY ARMED MEN”
6	PERPETRATOR: ID OF ORG(S)	“THE EXTRADITABLES”
7	PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: “THE EXTRADITABLES”
8	PHYSICAL TARGET: ID(S)	*
9	PHYSICAL TARGET: TOTAL NUM	*
10	PHYSICAL TARGET: TYPE(S)	*
11	HUMAN TARGET: ID(S)	“FEDERICO ESTRADA VELEZ” (“LIBERAL SENATOR”)
12	HUMAN TARGET: TOTAL NUM	1
13	HUMAN TARGET: TYPE(S)	GOVERNMENT OFFICIAL: “FEDERICO ESTRADA VELEZ”
14	TARGET: FOREIGN NATION(S)	–
15	INSTRUMENT: TYPE(S)	*
16	LOCATION OF INCIDENT	COLOMBIA: MEDELLIN (CITY)
17	EFFECT ON PHYSICAL TARGETS	*
18	EFFECT ON HUMAN TARGETS	*

Source Text (MUC-3) 7

TST-1-MUC3-0080
BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) - [REPORT] [JORGE ALONSO SIERRA VALENCIA] [TEXT] LIBERAL SENATOR FEDERICO ESTRADA VELEZ WAS KIDNAPPED ON 3 APRIL AT THE CORNER OF 60TH AND 48TH STREETS IN WESTERN MEDELLIN, ONLY 100 METERS FROM A METROPOLITAN POLICE CAI [IMMEDIATE ATTENTION CENTER]. THE ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER HAD LEFT HIS HOUSE WITHOUT ANY BODYGUARDS ONLY MINUTES EARLIER. AS WE WAITED FOR THE TRAFFIC LIGHT TO CHANGE, THREE HEAVILY ARMED MEN FORCED HIM TO GET OUT OF HIS CAR AND INTO A BLUE RENAULT.
HOURS LATER, THROUGH ANONYMOUS TELEPHONE CALLS TO THE METROPOLITAN POLICE AND TO THE MEDIA, THE EXTRADITABLES CLAIMED RESPONSIBILITY FOR THE KIDNAPPING. IN THE CALLS, THEY ANNOUNCED THAT THEY WILL RELEASE THE SENATOR WITH A NEW MESSAGE FOR THE NATIONAL GOVERNMENT. LAST WEEK, FEDERICO ESTRADA HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.

Text Example (MUC-5) 8

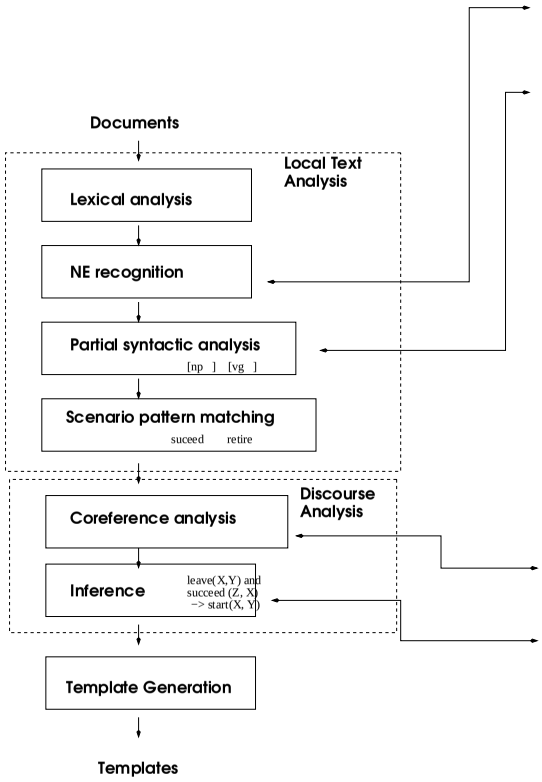
<DOC>
<DOCNO> 0592 </DOCNO>
<DD> NOVEMBER 24, 1989, FRIDAY </DD>
<SO>Copyright (c) 1989 Jiji Press Ltd.;</SO>
<TXT>
BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A JOINT VENTURE IN TAIWAN WITH A LOCAL CONCERN AND A JAPANESE TRADING HOUSE TO PRODUCE GOLF CLUBS TO BE SHIPPED TO JAPAN.
THE JOINT VENTURE, 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. THE MONTHLY OUTPUT WILL BE LATER RAISED TO 55,000 UNITS, BRIDGESTON SPORTS OFFICIALS SAID.
THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY BRIDGESTONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS SAID.
BRIDGESTONE SPORTS HAS SO FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUBS PARTS WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES.
WITH THE ESTABLISHMENT OF THE TAIWAN UNIT, THE JAPANESE SPORTS GOODS MAKER PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN JAPAN.
</TXT>
</DOC>

Template Example (MUC-5)

<TEMPLATE-0592-1> :=
DOC NR: 0592
DOC DATE: 241189
DOCUMENT SOURCE: "Jiji Press Ltd."
CONTENT: <TIE_UP_RELATIONSHIP-0592-1>
<TIE_UP_RELATIONSHIP-0592-1>:=
TIE-UP STATUS: EXISTING
ENTITY:<ENTITY-0592-1>
<ENTITY-0592-2>
<ENTITY-0592-3>
JOINT VENTURE CO:<ENTITY-0592-4>
OWNERSHIP: <OWNERSHIP-0592-1>
ACTIVITY:<ACTIVITY-0592-1>
<ENTITY-0592-1>:=
NAME: BRIDGESTONE SPORTS CO
ALIASES: "BRIDGESTONE SPORTS"
"BRIDGESTON SPORTS"
NATIONALITY: Japan (COUNTRY)
TYPE: COMPANY
ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-2>:=
NAME: UNION PRECISION CASTING CO
ALIASES: "UNION PRECISION CASTING"
"BRIDGESTON SPORTS"
LOCATION: Taiwan (COUNTRY)
NATIONALITY: Taiwan (COUNTRY)
TYPE: COMPANY
ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>
<ENTITY-0592-3>:=
NAME: TAGA CO
NATIONALITY: Japan (COUNTRY)
TYPE: COMPANY
ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>

<ENTITY-0592-4>:=
NAME: BRIDGESTONE SPORTS TAIWAN CO
ALIASES: "UNION PRECISION CASTING"
"BRIDGESTON SPORTS"
LOCATION: "KAOHSIUNG" (UNKNOWN) Taiwan (COUNTRY)
TYPE: COMPANY
ENTITY RELATIONSHIP:<ENTITY_RELATIONSHIP-0592-1>
<INDUSTRY-0592-1>:=
INDUSTRY-TYPE: PRODUCTION
PRODUCT/SERVICE: (CODE 39 "20,000 IRON AND 'METAL WOOD")
([CLUBS]")
<ENTITY_RELATIONSHIP-0592-1>:=
ENTITY1: <ENTITY-0592-1
<ENTITY-0592-2
<ENTITY-0592-3
ENTITY2: <ENTITY-0592-4
REL OF ENTITY2 TO ENTITY1: CHILD
STATUS: CURRENT
<ACTIVITY-0592-1>:=
INDUSTRY: <INDUSTRY-0592-1>
ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0592-4>)
START TIME: <TIME-0592-1>
<TIME-0592-1>:=
DURING: 0190
<OWNERSHIP-0592-1>:=
OWNED: <ENTITY-0592-4>
TOTAL-CAPITALIZATION: 20000000 TWD
OWNERSHIP-%: (<ENTITY-0592-3> 10)
(<ENTITY-0592-2> 15)
(<ENTITY-0592-1> 75)

NYU's IE system



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

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

[np: e1 Sam Schwartz (person)] [vg retired] as [np: e2 executive vice president] of [np: e3 the famous hot dog manufacturer], [np:e4 Hupplewhite Inc (organisation)]. [np: e5 He] [vg will be succeeded] by [np: e6 Harry Himmelfarb(person)].

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

e2	type:position	value:"executive vice president" company:e3
e3 = e4		
e7	leave-job	person:e1 position:e2
e8	succeed	person1:e6 person2:e5
e5 = e1		
e9	start-job	person:e6 position e2

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.

Named Entity recognition – as defined at MUC-6

- NE types:
  - ENAMEX (type= person, organisation, location)
  - TIMEX (type= time, date)
  - NUMEX (type= money, percent)
- Allowed to use gazetteers (fixed list containing names of a certain type, e.g. countries, last names, titles, state names, rivers...)
- ENAMEX is harder, more context dependent than TIMEX and NUMEX:
  - Is Granada a COMPANY or a LOCATION?
  - Is Washington a PERSON or a LOCATION?
  - Is Arthur Anderson a PERSON or an ORGANISATION?

Named Entity recognition – common approaches

- NE markup with subtypes:

  <ENAMEX TYPE='PERSON'>Flavel Donne</ENAMEX> is an analyst with <ENAMEX TYPE='ORGANIZATION'>General Trends</ENAMEX>, which has been based in <ENAMEX TYPE='LOCATION'>Little Spring</ENAMEX> since <TIMEX>July 1998</TIMEX>.
- Most systems use manually written regular expressions
  - Rules about mid initials, postfixes, titles
  - Gazetteers of common first names
  - Acronyms: Hewlett Packard Inc. → HP

PATTERN: "president of <company>" matches

executive vice president of Hupplewhite

Person names – evidence against gazetteers

- Gazetteer of full names impossible and not useful, as both first and last names can occur on their own
- Last name gazetteer impractical
  - Almost infinite set of name patterns possible: last names are productive (1.5M surnames in US alone)
  - Overlap with common nouns/verbs/adjectives
    - First 2 pages of Cambridge phone book include 237 names
    - Of those, 6 (2.5%) are common nouns: Abbey, Abbot, Acres, Afford, Airs, Alabaster
- First name gazetteer less impractical, but still not foolproof
  - First names can be surprising, eg. MUC-7 walk-through example: "Llennel Evangelista"
  - First names are productive, eg. Moonunit Zappa, Apple Paltrow . . .

Person Names – evidence against gazetteers

- – Overlap with common nouns:
  - River and Rain Phoenix, Moon Unit Zappa, Apple Paltrow
  - "Virtue names": Grace (134), Joy (390), Charity (480), Chastity (983), Constance, Destiny
  - "Month names": June, April, May
  - "Flower names": Rose, Daisy, Lily, Erica, Iris . . .
  - From US Social Security Administration's list of most popular girls' names in 1990, with rank:
    Amber (16), Crystal (41), Jordan (59), Jade (224), Summer (291), Ruby (300), Diamond (450), Infant (455), Precious (472), Genesis (528), Paris (573), Princess (771), Heaven (902), Baby (924) . . .
- Additional problem: non-English names alliterated into English; variant spellings
- Complicated name patterns with titles: Sammy Davis Jr,
  HRH The Prince of Wales, Dr. John T. Maxwell III

Name type ambiguity: more evidence against gazetteers

- Ambiguity of name types: *Columbia* (Org.) vs. (British) *Columbia* (Location) vs. *Columbia* (Space shuttle)
- Company names often use common nouns (“Next”, “Boots”, “Thinking Machines”. . .) and can occur in variations (“Peter Stuyvesant”, “Stuyvesant’)
- Coordination problems/ left boundary problems:
  - One or two entities in *China International Trust and Investment Corp invests $2m in. . .* ?
  - Unknown word at beginning of potential name: in or out?
    Suspended Ceilings Inc vs *Yesterday Ceilings Inc*
    Mason, Daily and Partners vs. *Unfortunately, Daily and Partners*
- Experiments show: simple gazetteers fine for locations (90%P/80%R) but not for person and organisations (80%P/50%R)

Mikheev et al. (1998): Cascading NE

- Staged combination of rule-based system with probabilistic partial matching
- Use machine learning to decide type of NE
- Use internal phrase structure of name
- Make high-precision decisions first
- Keep off decision about unsure items until all evidence has been seen
- Assume: one name type per discourse (article)
  - unless signalled by writer with additional context information

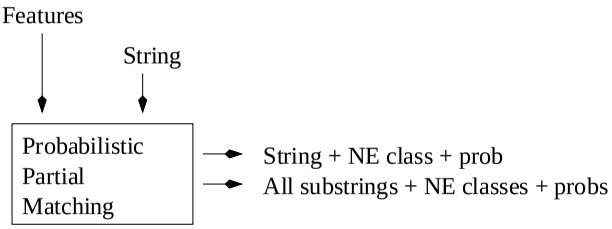
Mikheev et al. (1998): Examples of sure fire rules

Rule	Assign	Example
`(Xx+)+ (is|,) a? JJ* PROF`	PERS	Yuri Gromov, a former director
`(Xx+)+ is? a? JJ* REL`	PERS	John White is beloved brother
`(Xx+)+ himself`	PERS	White himself
`(Xx+)+, DD+ ,`	PERS	White, 33,
`share in (Xx+)+`	ORG	shares in Trinity Motors
`(Xx+)+ Inc.`	ORG	Hummingbird Inc.
`PROF (of|at|with) (Xx+)+`	ORG	director of Trinity Motors
`(Xx+)+ (region|area)`	LOC	Lower Beribidjan area

Mikheev et al. (1998): ML; External and internal evidence

External:

- Position in sentence (sentence initial)
- Word exists in lexicon in lowercase
- Word seen in lowercase in document



Internal:

- Contains any non-alpha characters
- Number of words it consists of
- Suffix, Prefix
- Adjectives ending in “an” or “ese” + whose root is in Gazetteer

Mikheev et al. (1998): Algorithm

1. Apply Grammar Rule Set 1 (“Sure fire” rules)
   → tag as definite NEs of given type
2. Use ML for variants (probabilistic partial match)
   - Generate all possible substrings of sure-fire tagged NEs:
     – *Adam Kluver Ltd* → *Adam Kluver*, *Adam Ltd*, *Kluver Ltd*
   - ME model gives probability for possible string and NE type
   - Tag all occurrences of NE in text (over prob. threshold) with type
3. Apply Grammar Rule Set 2 (Relaxed rules)
   - Mark anything that looks like a PERSON (using name grammar)
   - Resolve coordination, genitives, sentence initial capitalized modifiers
     – Coordinated or possessive name parts, or rest of sentence initial coordinated name seen on their own? If not, assume one name (*Murdoch’s News Corp, Daily, Bridge and Mason*)
4. Apply ML again (for new variants)
   - X and Y are of same type → resolved typo ‘‘`Un7ited States and Russia`’’
5. Apply specialised ME model to title (capitalisation, different syntax).

Mikheev et al – example text

MURDOCH SATELLITE CRASH UNDER FBI INVESTIGATION

London and Tomsk. The crash of Rupert Murdoch Inc’s news satellite yesterday is now under investigation by Murdoch and by the Sibirian state police. Clarity J. White, vice president of Hot Start, the company which produced the satellite’s ignition system, yesterday stated that her company considered human failure the most likely cause of the crash. Investigator Robin Black, 33, who investigates the crash for the FBI, recently arrived by train at the crash site in the Tomsk region. Neither White nor Black were available for comment today; Murdoch have announced a press conference for tomorrow.

Mikheev et al – potential NE hypotheses (all incorrect) 21

`LONDON and TOMSK`	Org
`Rupert Murdoch`	Person
`Murdoch`	Person
`Neither White`	Person
`Investigator Robin Black`	Person

Additional problem: `Black` and `White` have last names which overlap with adjectives and first names which overlap with common nouns (`Robin` and `Clarity`), thus they cannot be in a gazetteer.

Mikheev et al – After Step 1 22

MURDOCH SATELLITE CRASH UNDER FBI INVESTIGATION

London and Tomsk. The crash of Rupert Murdoch Inc(ORG)'s news satellite yesterday is now under investigation by Murdoch and by the Sibirian state police. Clarity J. White, vice president of Hot Start(ORG), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure the most likely cause of the crash. Investigator Robin Black(PERSON), 33, who investigates the crash for the FBI, recently arrived by train at the crash site in the Tomsk(LOC) region. Neither White nor Black were available for comment today; Murdoch have announced a press conference for tomorrow.

Underlined instances: newly suggested in this round

- Sure fire rules applied
- But exact extend of name not known yet: Investigator Robin Black? Black?

Mikheev et al – After Step 2 (Partial Match) 23

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London and Tomsk(LOC?). The crash of Rupert Murdoch Inc(ORG?)'s news satellite yesterday is now under investigation by Murdoch(ORG?) and by the Sibirian state police. Clarity J. White, vice president of Hot Start(ORG?), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure the most likely cause of the crash. Investigator Robin Black(PERSON?), 33, recently arrived by train at the crash site in the Tomsk(LOC?) region. Neither White nor Black(PERSON?) were available for comment today; Murdoch(ORG?) have announced a press conference for tomorrow.

Green instances: around from last round

- All instances from last round and their substrings are now hypothesized; they and their context are now subjected to ML

Mikheev et al – After Step 2 (Probabilistic Match) 24

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London and Tomsk(LOC$\surd$). The crash of Rupert Murdoch Inc(ORG$\surd$)'s news satellite yesterday is now under investigation by Murdoch(ORG$\surd$) and by the Sibirian state police. Clarity J. White, vice president of Hot Start(ORG$\surd$), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure as the most likely cause of the crash. Investigator Robin Black(PERS$\surd$), 33, recently arrived by train at the crash site in the Tomsk(LOC$\surd$) region. Neither White nor Black(PERS$\surd$) were available for comment today; Murdoch(ORG$\surd$) have announced a press conference for tomorrow.

- ML has reconfirmed some instances (Robin Black) and discarded others (Investigator Robin Black)

Mikheev et al – After Step 3

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London and Tomsk(LOC). The crash of Rupert Murdoch Inc(ORG)'s news satellite yesterday is now under investigation by Murdoch(ORG) and by the Sibirian state police. Clarity J. White(PERS?), vice president of Hot Start(ORG), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure as the most likely cause of the crash. Investigator Robin Black(PERS), 33, recently arrived by train at the crash site in the Tomsk(LOC) region. Neither White(PERS?) nor Black(PERS) were available for comment today; Murdoch(ORG) have announced a press conference for tomorrow.

- Relaxed rules: Mark everything as a possibility which roughly follows Name shape (blue, underlined)
- (plus confirmed NEs from last round in green)

Mikheev et al – After Step 4

MURDOCH SATELLITE CRASH UNDER INVESTIGATION

London(LOC$\surd$) and Tomsk(LOC). The crash of Rupert Murdoch Inc(ORG)'s news satellite yesterday is now under investigation by Murdoch(ORG) and by the Sibirian state police. Clarity J. White(PERS$\surd$), vice president of Hot Start(ORG), the company which produced the satellite's ignition system, yesterday stated that her company considered human failure as the most likely cause of the crash. Investigator Robin Black(PERS), 33, recently arrived by train at the crash site in the Tomsk(LOC) region. Neither White(PERS$\surd$) nor Black(PERS) were available for comment today; Murdoch(ORG) have announced a press conference for tomorrow.

- Some of these possibilities reconfirmed by ML, others discarded
- "London" found by X and Y rule.
- Missing step: different segmentation and ML for title; 'Murdoch' is found there.

Mikheev et al: Results in MUC-7

93.39% combined P and R – best and statistically different from next contender

		ORG		PERSON		LOC	
		R	P	R	P	R	P
1	Sure fire rules	42	98	40	99	36	96
2	Partial Match 1	75	98	80	99	69	93
3	Relaxed Rules	83	96	90	98	86	93
4	Partial Match 2	85	96	93	97	88	93
5	Title Assignment	91	95	95	97	95	93

- System design: Keep precision high at all stages, raise recall if possible
- Gazetteers improve performance, but system can determine persons and organizations reasonably well even without any gazetteer (ORG: P86/R85; PERSON: P90/R95), but not locations (P46/R59)

Summary of today

- IE consists of different tasks (as defined by MUC): NE, CO, TE, ST
- Today: NE
  - Principal problems with NE
  - NE with manual rules
  - Mikheev et al. (1998)
    - Use internal and external evidence
    - Cascaded design: commit in order of confidence/supportive evidence from text, not in text order!

- Mikheev, Moens and Grover (1998). Description of the LTG system. MUC-7 Proceedings.
- Mikheev, Moens and Grover (1999). Named Entity Recognition without Gazetteers. EACL'99
- R. Grishman (1997): Information Extraction: Techniques and challenges, in: Information Extraction, Springer Verlag, 1997.

Information Retrieval

Lecture 6: Information Extraction and Bootstrapping

Computer Science Tripos Part II

UNIVERSITY OF CAMBRIDGE

Stephen Clark

Natural Language and Information Processing (NLIP) Group

sc609@cl.cam.ac.uk

Last time

- Range of problems that make named entity recognition (NE) hard
- Mikheev et al's (1998) cascading NE system
- NE is the simplest kind of IE task: no relations between entities must be determined
- NIST MUC conferences pose three kinds of harder IE tasks
- Today: more of the full task (scenario templates), and on learning

Lexico-semantic patterns

- "Flattened-out" semantic representations with lexemes directly hard-wired into them
- String-based matching with type of semantic category to be found directly expressed in lexical pattern
- Problem with all string-based mechanisms: generalisation to other strings with similar semantics, and to only those
- Do generalisation by hand...
  - <Perpetrator> (APPOSITION) {blows/blew/has blown} {himself/herself} up
  - <Perpetrator> detonates
  - {blown up/detonated} by <Perpetrator>
- Manual production of patterns is time-consuming, brittle, and not portable across domains

Learning of lexico-semantic patterns (Riloff 1993)

- UMASS participant system in MUC-4: AutoSlog
- Lexico-semantic patterns for MUC-3 took 1500 person hours to build → knowledge engineering bottleneck
- AutoSlog achieved 98% performance of manual system; AutoSlog dictionary took 5 person hours to build
- "Template mining:"
  - Use MUC training corpus (1500 texts + human answer keys; 50% non-relevant texts) to learn contexts
  - Have human check the resulting templates (30% - 70% retained)

Lexico-syntactic-semantic patterns (Riloff 1993)

- 389 Patterns ("concept nodes") with enabling syntactic conditions, e.g. active or passive:
  - kidnap-passive: <VICTIM> expected to be subject
  - kidnap-active: <PERPETRATOR> expected to be subject
- Hard and soft constraints for fillers of slots
  - Hard constraints: selectional restrictions; soft constraints: semantic preferences
- Semantic lexicon with 5436 entries (including semantic features)

Heuristics for supervised template mining (Riloff 1993)

- Stylistic conventions: relationship between entity and event made explicit in first reference to the entity
- Find key word there which triggers the pattern: *kidnap, shot,*
- Heuristics to find these trigger words
- Given: filled template plus raw text. Algorithm:
  - Find first sentence that contains slot filler
  - Suggest good conceptual anchor point (trigger word)
  - Suggest a set of enabling conditions

"the diplomat was kidnapped" + VICTIM: the diplomat Suggest: <SUBJECT> passive-verb + trigger=kidnap

Learning of lexico-semantic patterns (Riloff 1993)

System uses 13 "heuristics" (= syntactic patterns):

EXAMPLE	PATTERN
<victim> was murdered	<subject> passive-verb
<perpetrator> bombed	<subject> active-verb
<perpetrator> attempted to kill	<subject> verb infinitive
<victim> was victim	subject auxiliary <noun>
killed <victim>	passive-verb <dobj>
bombed <target>	active-verb <dobj>
to kill <victim>	infinitive <dobj>
threatened to attack <target>	verb infinitive <dobj>
killing <victim>	gerund <dobj>
fatality was <victim>	noun auxiliary <dobj>
bomb against <target>	noun prep <np>
killed with <instrument>	active-verb prep <np>
was aimed at <target>	passive-verb prep <np>

Riloff 1993: a good concept node

ID: DEV-MUC4-0657
Slot Filler: “public buildings”
Sentence: IN LA OROYA, JUNIN DEPARTMENT, IN THE CENTRAL PERUVIAN MOUNTAIN RANGE, PUBLIC BUILDINGS WERE BOMBED AND A CAR-BOMB WAS DETONATED.

CONCEPT NODE	
Name:	target-subject-passive-verb-bombed
Trigger:	bombed
Variable slots:	(target (*S* 1))
Constraints:	(class phys-target *S*)
Constant slots:	(type bombing)
Enabling Conditions:	((passive))

Riloff 1993: another good concept node

ID: DEV-MUC4-0071
Slot Filler: “guerrillas
Sentence: THE SALVADORAN GUERRILLAS ON MAR_12_89, TODAY, THREATENED TO MURDER INDIVIDUALS INVOLVED IN THE MAR_19_88 PRESIDENTIAL ELECTIONS IF THEY DO NOT RESIGN FROM THEIR POSTS.

CONCEPT NODE	
Name:	perpetrator-subject-verb-infinitive-threatened-to-murder
Trigger:	murder
Variable slots:	(perpetrator (*S* 1))
Constraints:	(class perpetrator *S*)
Constant slots:	(type perpetrator)
Enabling Conditions:	((active) (trigger-preceded-by? ’to ’threatened))

Riloff 1993: a bad concept node

ID: DEV-MUC4-1192
Slot Filler: “gilberto molasco
Sentence: THEY TOOK 2-YEAR-OLD GILBERTO MOLASCO, SON OF PATRICIO RODRIGUEZ, AND 17-OLD ANDRES ALGUETA, SON OF EMIMESTO ARGUETA.

CONCEPT NODE	
Name:	victim-active-verb-dobj-took
Trigger:	took
Variable slots:	(victim (*DOBJ* 1))
Constraints:	(class victim *DOBJ*)
Constant slots:	(type kidnapping)
Enabling Conditions:	((active))

Riloff 1993: evaluation

System/Test Set	Recall	Prec	F-measure
MUC-4/TST3	46	56	50.5
AutoSlog/TST3	43	56	48.7
MUC-4/TST4	44	40	41.9
AutoSlog/TST4	39	45	41.8

- 5 hours of sifting through AutoSlog’s patterns
- Porting to new domain in less than 10 hours of human interaction
- But: creation of training corpus ignored in this calculation

Agichtein, Gravano (2000): Snowball

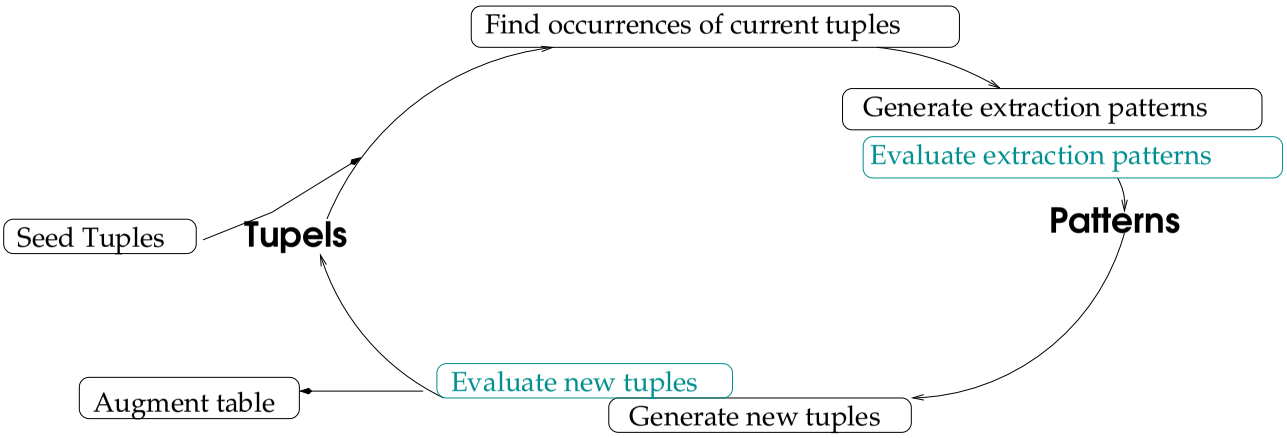
- Find locations of headquarters of a company and the corresponding company name ($< o, l >$ tuples)

Organisation	Location of Headquarters
Microsoft	Redmond
Exxon	Irving
IBM	Armonk
Boeing	Seattle
Intel	Santa Clara

  "Computer servers at Microsoft's headquarters in Redmond"

- Use minimal human interaction (handful of positive examples)
  - no manually crafted patterns
  - no large annotated corpus (IMass system at MUC-6)
- Automatically learn extraction patterns
- Less important to find every occurrence of patterns; only need to fill table with confidence

Agichtein, Gravano (2000): Bootstrapping



Agichtein, Gravano (2000): Overall process

- Start from table containing some $< o, l >$ tuples (which must exist in document collection)
- Perform NE (advantage over prior system DIPRE (Brin 98))
- System searches for occurrences of the example $< o, l >$ tuples in documents
- System learns extraction patterns from these example contexts, e.g.:

<ORGANIZATION> 's headquarters in <LOCATION>
<LOCATION>-based <ORGANIZATION>

- Evaluate patterns; use best ones to find new $< o, l >$ tuples
- Evaluate new tuples, choose most reliable ones as new seed tuples
- Iteratively repeat the process

Agichtein, Gravano (2000): Context generalisation and patterns

A SNOWBALL pattern is a 5-tuple <left,tag1,middle,tag2,right>

left	Tag1	middle	Tag2	right
The	Irving	-based	Exxon Corporation	
<{<the, 0.2>},	LOCATION,	{<-,0.5> <based, 0.5>},	ORGANIZATION,	{} >

- Associate term weights as a function of frequency of term in context
- Normalize each vector so that norm is 1; then multipy with weights $W_{left}, W_{right}, W_{mid}$.
- Degree of match between two patterns $t_p =< l_p, t_1, m_p, t_2, r_p >$ and $t_s =< l_s, t'_1, m_s, t'_2, r_s >$:

$$match(t_p, t_s) = l_p l_s + m_p m_s + r_p r_s \text{ (if tags match, 0 otherwise)}$$

Agichtein, Gravano (2000): Pattern generation

- Similar contexts form a pattern
  - Cluster vectors using a clustering algorithm (minimum similarity threshold τ_{sim})
  - Vectors represented as cluster centroids $\bar{l}_s, \bar{m}_s, \bar{r}_s$
- Generalised Snowball pattern defined via centroids:

$$< \bar{l}_s, tag_1, \bar{m}_s, tag_2, \bar{r}_s >$$

- Remember for each Generalised Snowball pattern
  - All contexts it came from
  - The distances of contexts from centroid

Agichtein, Gravano (2000): Productivity/Reliability

- We want productive and reliable patterns
  - productive but not reliable:
    $< \{\}, ORGANIZATION, \{<'',', 1>\}, LOCATION, \{\} >$

    "Intel, Santa Clara, announced that..."
    "Invest in Microsoft, New York-based analyst Jane Smith said..."
  - reliable but not productive:
    $< \{\}, ORGANIZATION, \{< whose, 0.1 >, < headquarter, 0.4 >, < is, 0.1 >< located, 0.3 >, < in, 0.09 >, < nearby, 0.01 >\}, LOCATION, \{\} >$

    "Exxon, whose headquarter is located in nearby Irving..."
- Eliminate patterns supported by less than $\tau_{sup} < o, l >$ tuples

Agichtein, Gravano (2000): Pattern reliability

- If P predicts tuple $t =< o, l >$ and there is already tuple $t' =< o, l' >$ with high confidence, then: if $l = l' \rightarrow P.positive$++, otherwise $P.negative$++ (uniqueness constraints: organization is key).
- Pattern reliability: $Conf(P) = \frac{P.positive}{P.positive+P.negative}$ (range [0..1])
- Example:
  $P_{43} =< \{\}, ORGANIZATION, \{<'',', 1>\}, LOCATION, \{\} >$ matches
  1. Exxon, Irving, said... (CORRECT: in table)
  2. Intel, Santa Clara, cut prices (CORRECT: in table)
  3. invest in Microsoft, New York-based analyst (INCORRECT, contradicted by entry <Microsoft, Redmont>)
  4. found at ASDA, Irving. (????, unknown, no contradiction → disregard evidence)
- disregard unclear evidence such as 4.
- Thus, $Conf(P_{43}) = \frac{2}{2+1}$

Agichtein, Gravano (2000): Pattern confidence

- Consider productivity, not just reliability:

$$Conf_{RlogF}(P) = Conf(P) log_2(P.positive)$$

- Normalized $Conf_{RlogFNorm}(P)$:

$$Conf_{RlogFNorm}(P) = \frac{Conf_{RlogF}(P)}{max_{i \in \mathcal{P}} Conf(i)}$$

  (this brings $Conf_{RlogFNorm}(P)$ into range [0...1])
- $max_{i \in \mathcal{P}} Conf(i)$ is the largest confidence value seen with any pattern
- $Conf_{RlogFNorm}(P)$ is a rough estimate of the probability of pattern P producing a valid tuple (called $Conf(P)$ hereafter)

Agichtein, Gravano (2000): Tuple evaluation I

- Confidence of a tuple T is probability that at least one valid tuple is produced:

$$Conf(T) = 1 - \prod_{i=0}^{|P|}(1 - Conf(P_i)Match(C_i, P_i))$$

$P = \{P_i\}$ is the set of patterns that generated T
C_i is the context associated with an occurrence of T
$Match(C_i, P_i)$ is goodness of match between P_i and C_i

- Explanation: probability of every pattern matched incorrectly:

$$Prob(T \text{ is NOT valid}) = \prod_{i=0}^{|P|}(1 - P(i))$$

- Formula due to the assumption that for an extracted tuple T to be valid, it is sufficient that at least one pattern matched the “correct” text context of T.

Agichtein, Gravano (2000): Tuple evaluation II

- Then reset confidence of patterns:

$$Conf(P) = Conf_{new}(P)W_{updt} + Conf_{old}(P)(1 - W_{updt})$$

W_{updt} controls learning rate: does system trust old or new occurrences more? Here: $W_{updt} = 0.5$

- Throw away tuples with confidence $< \tau_t$

Agichtein, Gravano (2000): Results

Conf	middle	right
1	<based, .53>, <in, .53>	<”,” ,.01>
.69	<””, .42>,<s, .42>,<headquarters, .42>,<in,.42>	
.61	<(,.93>	<),.12>

- Use training corpus to set parameters: $\tau_{sim}, \tau_t, \tau_{sup}$, I_{max}, W_{left}, W_{right}, W_{middle}
- Only input: 5 $< o, l >$ tuples
- Punctuation matters: performance decreases when punctuation is removed
- Recall b/w .78 and .87 ($\tau_{sup} > 5$); precision .90 ($\tau_{sup} > > 4$)
- High precision possible (.96 with τ_t = .8); remaining problems come from NE recognition
- Pattern evaluation step responsible for most improvement over DIPRE

Summary: IE and template matching, learning

- Possible to learn simple relations from positive examples (Snowball)
- Possible to learn more diverse relations from annotated training corpus (Riloff)
- Even modest performance can be useful
  - Later manual verification
  - In circumstances where there would be no time to review source documents, so incomplete extracted information is better than none

Summary: IE Performance

Current methods perform well if

- Information to be extracted is expressed directly (no complex inference is required)
- Information is predominantly expressed in a relatively small number of forms
- Information is expressed locally within the text

Difference between IE and QA (next time):

- IE is domain dependent, open-domain QA is not

Literature

- Ellen Riloff, Automatically constructing a dictionary for information extraction tasks. In Proc. 11th Ann. Conference of Artificial Intelligence, p 811-816, 1993
- Eugene Agichtein, Luis Gravano: Snowball: Extracting Relations from Large Plain-Text Collections, Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000

Information Retrieval

Lecture 7: Question Answering

Computer Science Tripos Part II

Stephen Clark

Natural Language and Information Processing (NLIP) Group

sc609@cl.cam.ac.uk

Question Answering: Task definition in TREC-QA

- QA Track since TREC-1999: Open-domain factual textual QA
- Task requirements (in comparison with IR):
  1. Input: NL questions, not keyword-based queries
  2. Output: answers, not documents
- Rules:
  - All runs completely automatic
  - Frozen systems once questions received; answers back to TREC within one week
  - Answers may be extracted or automatically generated from material in document collection only
  - The use of external resources (dictionaries, ontologies, WWW) is allowed
  - Each returned answer is checked manually by TREC-QA (no comparison to gold standard)

TREC QA: Example questions

TREC-8	How many calories are there in a Big Mac? Where is the Taj Mahal?
TREC-9	Who invented the paper clip? How much folic acid should an expectant mother take daily? Who is Colin Powell?
TREC-10	What is an atom? How much does the human adult female brain weigh? When did Hawaii become a state?

Questions in TREC

- Type of question: reason, definition, list of instances, context-sensitive to previous questions (TREC-10)
- Source of question: invented for evaluation (TREC-8); since TREC-9 mined from logs (Encarta, Excite)
  - → strong impact on task: more realistic questions are harder on assessors and systems, but more representative for training
- Type of answer string: 250 Bytes (TREC-8/9, since TREC-12); 50 Bytes (TREC-8–10); exact since TREC-11
- Guarantee of existence of answer: no longer given since TREC-10

Examples of answer strings

What river in the US is known as the Big Muddy?

System A:	the Mississippi
System B:	Known as Big Muddy, the Mississippi is the longest
System C:	as Big Muddy , the Mississippi is the longest
System D:	messed with . Known as Big Muddy , the Mississip
System E:	Mississippi is the longest river in the US
System F:	the Mississippi is the longest river in the US
System G:	the Mississippi is the longest river(Mississippi)
System H:	has brought the Mississippi to its lowest
System I:	ipes.In Life on the Mississippi,Mark Twain wrote t
System K:	Southeast;Mississippi;Mark Twain;officials began
System L:	Known; Mississippi; US,; Minnessota; Cult Mexico
System M:	Mud Island,; Mississippi; “The; history; Memphis

Decreasing quality of answers

Manual checking of answers

- Systems return [docid, answer-string] pairs; mean answer pool per question judged: 309 pairs
- Answers judged in the context of the associated document
- ”Objectively” wrong answers okay if document supports them
  - Taj Mahal
- Considerable disagreement in terms of absolute evaluation metrics
- But relative MRRs (rankings) across systems very stable

Labels

- Ambiguous answers are judged as “incorrect”:

  What is the capital of the Kosovo?

  250B answer:

  protestors called for intervention to end the “Albanian uprising”. At Vucitrn, 20 miles northwest of Pristina, five demonstrators were reported injured, apparently in clashes with police. Violent clashes were also repo

- Answers need to be supported by the document context → the second answer is “unsupported”:

  What is the name of the late Phillippine President Marco’s wife?

  – Ferdinand Marcos and his wife Imelda. . . → [supported]

  – Imelda Marcos really liked shoes. . . → [unsupported]

List task (TREC-10, since TREC-12)

- 25 questions: retrieve a given target number of instances of something
- Goal: force systems to assemble an answer from multiple strings
  – `Name 4 US cities that have a ‘‘Shubert’’ theater`
  – `What are 9 novels written by John Updike?`
  – `What are six names of navigational satellites?`
  – `Name 20 countries that produce coffee.`
- List should not be easily located in reference work
- Instances are guaranteed to exist in collection
- Multiple documents needed to reach target, though single documents might have more than one instance
- Since TREC-12: target number no longer given; task is to find all

MRR: Mean reciprocal rank

- Task is precision-oriented: only look at top 5 answers
- Score for individual question i is the reciprocal rank r_i where the first correct answer appeared (0 if no correct answer in top 5 returns).

$$RR_i = \frac{1}{r_i}$$

- Possible reciprocal ranks per question: [0, 0.2, 0.25, 0.33, 0.5, 1]
- Score of a run (MRR) is mean over n questions:

$$MRR = \frac{1}{n} \sum_{i=1}^{n} RR_i$$

Example: Mean reciprocal rank


```
162:  What is the capital of Kosovo?
```

```
1  18 April, 1995, UK GMT Kosovo capital
2  Albanians say no to peace talks in Pr
3  0 miles west of Pristina, five demon
4  Kosovo is located in south and south
5  The provincial capital of the Kosovo
```

$\rightarrow RR_{162} = \frac{1}{3}$

```
23:  Who invented the paper clip?
```

```
1  embrace Johan Vaaler, as the true invento
2  seems puzzling that it was not invented e
3  paper clip.  Nobel invented many useful th
4  modern-shaped paper clip was patented in A
5  g Johan Valerand, leaping over Norway, in
```

$\rightarrow RR_{23} = 1$

```
2:  What was the monetary value of the Nobel Peace
Prize in 1989?
```

```
1  The Nobel poll is temporarily disabled.  1994 poll
2  perience and scientific reality, and applied to socie
3  Curies were awarded the Nobel Prize together with Beqc
4  the so-called beta-value.  $40,000 more than expected
5  that is much greater than the variation in mean value
```

$\rightarrow RR_2 = 0$

$\rightarrow MRR = \frac{\frac{4}{3}}{3} = .444$

Other QA evaluation metrics used in TREC

- Average accuracy since 2003: only one answer per question allowed; accuracy is $\frac{\textit{Answers correct}}{\textit{Total Answers}}$
- Confidence-weighted score: systems submit one answer per question and order them according to the confidence they have in the answer (with their best answer first in the file)

$$\frac{1}{Q}\sum_{i=1}^{Q}\frac{\#\textit{correct in first } i}{i}$$

(Q being the number of questions). This evaluation metric (which is similar to Mean Average Precision) was to reward systems for their confidence in their answers, as answers high up in the file participate in many calculations.

Results

- In TREC-8, 9, 10 best systems returned MMR of .65–.70 for 50B answers, answering around 70–80% of all questions
- In 55% of the cases where answer was found in the first 5 answers, this answer was in rank 1
- Accuracy of best system in TREC-10's list task had an accuracy of .75
- The best confidence-weighted score in TREC-11 achieved was .856 (NIL-prec .578, NIL recall .804)
- TREC-12 (exact task): Best performance was an accuracy of .700

QA systems

- Overview of three QA systems:
- Cymphony system (TREC-8)
  - NE plus answer type detection
  - Shallow parsing to analyse structure of questions
- SMU (TREC-9)
  - Matching of logical form
  - Feedback loops
- Microsoft (TREC-10)
  - Answer redundancy and answer harvesting
  - Claim: “Large amounts of data make intelligent processing unnecessary.”

Overall algorithm

- Question Processing
  - Shallow parse
  - Determine expected answer type
  - Question expansion
- Document Processing
  - Tokenise, POS-tag, NE-index
- Text Matcher (= Answer production)
  - Intersect search engine results with NE
  - Rank answers

Named entity recognition

- Over 80% of 200 TREC-8 questions ask for a named entity (NE)
- NE employed by most successful systems in TREC (Verhees and Tice, 2000)
- MUC NE types: person, organisation, location, time, date, money, percent
- Textract covers additional types:
  - frequency, duration, age
  - number, fraction, decimal, ordinal, math equation
  - weight, length, temperature, angle, area, capacity, speed, rate
  - address, email, phone, fax, telex, www
  - name (default proper name)
- Textract subclassifies known types:
  - organisation → company, government agency, school
  - person → military person, religious person

Expected answer type

Who won the 1998 Nobel Peace Prize?

Expected answer type:	PERSON
Key words:	won, 1998, Nobel, Peace, Prize

Why did David Koresh ask the FBI for a word processor?

Expected answer type:	REASON
Key words:	David, Koresh, ask, FBI, word, processor

Question Expansion:

Expected answer type:	[because | because of | due to | thanks to | since | in order to | to VP]
Key words:	[ask|asks|asked|asking, David, Koresh, FBI, word, processor]

FST rules for expected answer type

R1: Name NP(city | country | company) → CITY|COUNTRY|COMPANY

VG[name] NP[a country] that VG[is developing] NP[a magnetic levitation railway system]

R2: Name NP(person_w) → PERSON

VG[Name] NP[the first private citizen] VG[to fly] PP[in space]

("citizen" belongs to word class `person_w`).

R3: CATCH-ALL: proper noun

Name a film that has won the Golden Bear in the Berlin Film Festival.

Direct matching of question words

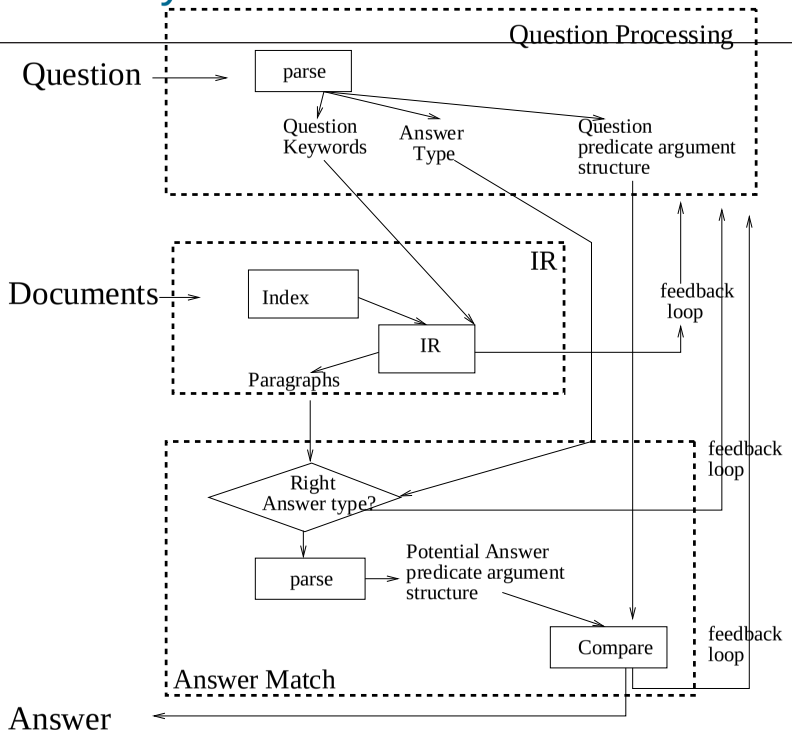
who/whom →	PERSON
when →	TIME/DATE
where/what place →	LOCATION
what time (of day) →	TIME
what day (of the week) →	DAY
what/which month →	MONTH
how often →	FREQUENCY
...	

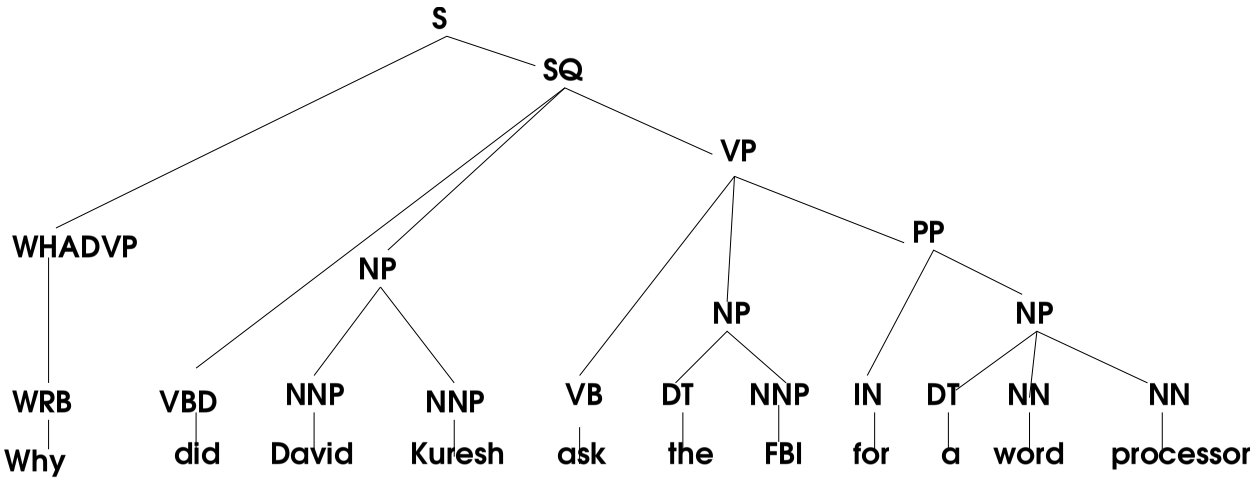
This classification happens only if the previous rule-based classification did not return unambiguous results.

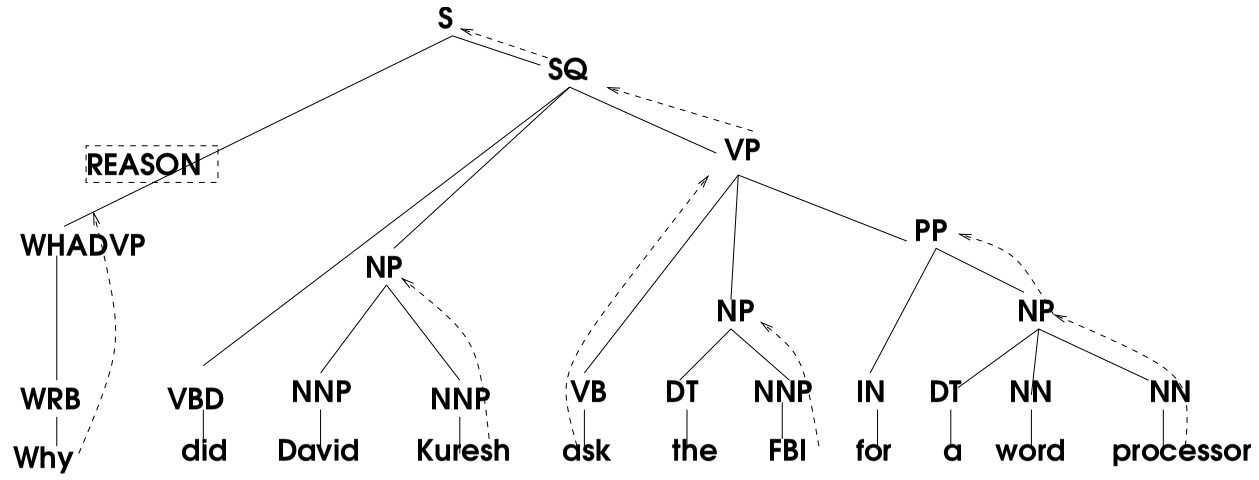
The Southern Methodist University (SMU) system (Harabagiu et al.)

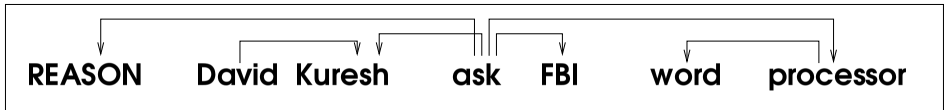
- Example of a deep processing system which has been extremely successful in TREC-QA (clear winner in most years)
- Machinery beyond answer type determination:
  1. Variants/feedback loops: morphological, lexical, syntactic, by reasoning
  2. Comparison between answer candidate and question on basis of logical form
- Deep processing serves to
  - capture semantics of open-domain questions
  - justify correctness of answers

Overview of SMU system




S
SQ
VP
WHADVP
NP
PP
NP
NP
WRB
VBD
NNP
NNP
VB
DT
NNP
IN
DT
NN
NN
Why
did
David
Kuresh
ask
the
FBI
for
a
word
processor


S
SQ
REASON
VP
WHADVP
PP
NP
NP
NP
WRB
VBD
NNP
NNP
VB
DT
NNP
IN
DT
NN
NN
Why
did
David
Kuresh
ask
the
FBI
for
a
word
processor


REASON
David
Kuresh
ask
FBI
word
processor

SMU: Variants ("Feedback loops")

- Morphological (+40%):
  - *Who invented the paper clip?* — Main verb "invent", ANSWER-TYPE "who" (subject) → add keyword "inventor"
- Lexical (+52%; used in 129 questions):
  - *How far is the moon?* — "far" is an attribute of "distance"
  - *Who killed Martin Luther King?* — "killer" = "assassin"
- Semantic alternations and paraphrases, abductive reasoning (+8%; used in 175 questions)
  - *How hot does the inside of an active volcano get?*
  - Answer in "lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit"
  - Facts needed in abductive chain:
    - volcano IS-A mountain; lava PART-OF volcano
- Combination of loops increases results considerably (+76%)

At the other end of the spectrum: the Microsoft system

- Circumvent difficult NLP problems by using more data
- The web has 2 billion indexed pages
- Claim: deep reasoning is only necessary if search ground is restricted
- The larger the search ground, the greater the chance of finding answers with a simple relationship between question string and answer string:

Who killed Abraham Lincoln?

DOC 1	John Wilkes Booth is perhaps America's most infamous assassin. He is best known for having fired the bullet that ended Abraham Lincoln's life.	TREC
DOC 2	John Wilkes Booth killed Abraham Lincoln.	web

The Microsoft system: Methods

1. Question processing is minimal: reordering of words, removal of question words, morphological variations
2. Matching done by Web query (google):
   - Extract potential answer strings from top 100 summaries returned
3. Answer generation is simplistic:
   - Weight answer strings (frequency, fit of match) – learned from TREC-9
   - Shuffle together answer strings
   - Back-projection into TREC corpus: keywords + answers to traditional IR engine
4. Improvement: Expected answer type filter (24% improvement)
   - No full-fledged named entity recognition

Query string generation

Rewrite module outputs a set of 3-tupels:

- Search string
- Position in text where answer is expected with respect to query string : LEFT|RIGHT|NULL
- Confidence score (quality of template)

Who is the world's richest man married to?

[+is the world's richest man married to	LEFT	5]
[the +is world's richest man married to	LEFT	5]
[the world's +is richest man married to	RIGHT	5]
[the world's richest +is man married to	RIGHT	5]
[the world's richest man +is married to	RIGHT	5]
[the world's richest man married +is to	RIGHT	5]
[the world's richest man married to +is	RIGHT	5]
[world's richest man married	NULL	2]
[world's AND richest AND married	NULL	1]

String weighting

- Obtain 1-grams, 2-grams, 3-grams from google short summaries
- Score each n-gram n according to the weight r_q of query q that retrieved it
- Sum weights across all summaries containing the ngram n (this set is called S_n)

$$w_n = \sum_{n \in S_n} r_q$$

w_n: weight of ngram n
S_n: set of all retrieved summaries which contain n
r_q: rewrite weight of query q

Answer string generation

- Merge similar answers (ABC + BCD $\rightarrow$ ABCD)
  - Assemble longer answers from answer fragments
  - Weight of new n-gram is maximum of constituent weights
  - Greedy algorithm, starting from top-scoring candidate
  - Stop when no further ngram tiles can be detected
  - But: cannot cluster “redwoods” and “redwood trees”
- Back-projection of answer
  - Send keywords + answers to traditional IR engine indexed over TREC documents
  - Report matching documents back as “support”
- Always return NIL on 5th position

The Microsoft system: Examples

- Time sensitivity of questions:
  Q1202: Who is the Governor of Alaska? → system returns governor in 2001, but TREC expects governor in 1989.
- Success stories:

Question	Answer	TREC document
What is the birth-stone for June?	Pearl	for two weeks during June (the pearl is the birth-stone for those born in that month)
What is the rainiest place on Earth?	Mount Wailaleale	and even Pago Pago, noted for its prodigious showers, gets only about 196 inches annually (The titleholder, according to the National Geographic Society, is Mount Wailaleale in Hawaii, where about 460 inches of rain falls each year).

Microsoft system: Discussion

- Results: mid-range (.347 MRR, 49% no answer)
- Development time of less than a month
- Produced “exact strings” before TREC-11 demanded it: average returned length 14.6 bytes
- Does this system undermine of QA as a gauge for NL understanding?
  - If TREC wants to measure straight performance on factual question task, less NLP might be needed than previously thought
  - But if TREC wants to use QA as test bed for text understanding, it might now be forced to ask “harder” questions
- And still: the really good systems are still the ones that do deep NLP processing!

Summary

- Open domain, factual question answering
- TREC: Source of questions matters (web logs v. introspection)
- Mean reciprocal rank main evaluation measure
- MRR of best systems 0.68 - 0.58
- Best systems answer about 75% of questions in the first 5 guesses, and get the correct answer at position 1.5 on avg ($\frac{1}{.66}$)
- System technology
  - NE plus answer type detection (Cymphony)
  - Matching of logical form, Feedback loops (SMU)
  - Answer redundancy and answer harvesting (Microsoft)

Literature

- Teufel (2007): Chapter *An Overview of evaluation methods in TREC Ad-hoc Information Retrieval and TREC Question Answering*. In: L. Dybkjaer, H. Hemsen, W. Minker (Eds.) Evaluation of Text and Speech Systems. Springer, Dordrecht, The Netherlands.
- Ellen Voorhees (1999): The TREC-8 Question Answering Track Report, Proceedings of TREC
- R. Srihari and W. Li (1999): “Information-extraction supported question answering”, TREC-8 Proceedings
- S. Harabagiu et al (2001), “The role of lexico-semantic feedback in open-domain textual question-answering”, ACL-2001
- E. Brill et al (2001), “Data intensive question answering”, TREC-10 Proceedings

Information Retrieval

Lecture 8: Automatic Summarisation

Computer Science Tripos Part II

Stephen Clark

Natural Language and Information Processing (NLIP) Group

sc609@cl.cam.ac.uk

Summarisation – an impossible task?

- Summarisation is intelligent and linguistically viable information compression
- Part of human activity in many different genres
  - TV guide: movie plot summaries
  - Blurb on back of book
  - Newsflashes
  - Subtitles
- Why do research in automatic summarisation?
  - Practical reasons: information compression needed in today's information world
  - Scientific reasons: summarisation is a test bed for current document understanding capabilities

Text Summarisation

- Compress the “most important” points of a text, express these main points in textual form
- Information reduction
- Different types of summaries
  - informative/indicative
    - informative: summary replaces full document v.
    - indicative: decision aid for question “should I read the full document?”
  - abstract/extract
    - abstract (generated text) v.
    - extract (verbatim text snippets)

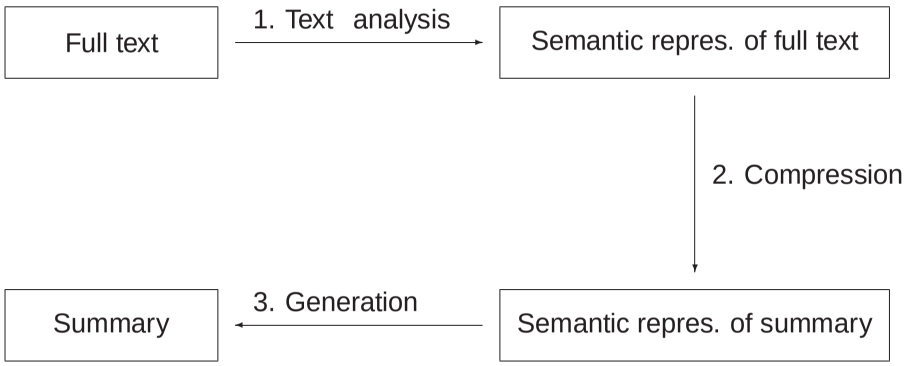
Properties of a good summary

- Considerably shorter than the input text
- Covers main points of input text
- Truth-preserving
- A good text in its own right (coherence...)
- Additional goals: flexibility (with respect to length, user, task)

Human abstracting

- Abstractors are employed at indexing/abstracting companies which produce abstract journals
- Need expert knowledge about summarising and about domain
- Several studies of human abstractors (Cremmins 1996, Endress-Niggemeyer 1995, Liddy 1991)
- Studies show that human abstractors
  - extract textual material, rework it (Cremmins, E-N)
  - only create new material from scratch when they have to, by generalisation and inference (Cremmins, E-N)
  - have a consistent building plan of a summary in their minds, but agree more on type of information to be put into summary than on the actual sentences (Liddy)
- But: Instructions for abstractors too abstract to be used for actual algorithms

Text summarisation: the deep model



Steps of the deep model:

1. Analysis of text into semantic representation
2. Manipulation (compression) of semantic representation
3. Text generation from semantic representation

- Compression methods exist (step 2)
  - Summarisation model by Kintsch and van Dijk (1979), based on propositions and human memory restrictions
  - Reasoning theories, e.g. by Lehnert (1982)
- Natural and flexible text generation exists (step 3), working from semantic representation
  - McKeown et al.: Generation from basketball game statistics, weather reports
  - Moore and DiEugenio: Generation of tutor's explanations
- Bottleneck: text analysis (step 1)

Summarisation by fact extraction (Radev and McKeown 1998, CL) 94

Compress several descriptions about the same event from multiple news stories

MESSAGE: ID	**TST-REU-0001**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 3, 1996 11:30
PRIMSOURCE: SOURCE	
INCIDENT: DATE	March 3, 1996
INCIDENT: LOCATION	Jerusalem
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: 18" "wounded: 10"
PERP: ORGANIZATION ID	

MESSAGE: ID	**TST-REU-0002**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 4, 1996 07:20
PRIMSOURCE: SOURCE	Israel Radio
INCIDENT: DATE	March 4, 1996
INCIDENT: LOCATION	Tel Aviv
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: at least 10" "wounded: 30"
PERP: ORGANIZATION ID	

MESSAGE: ID	**TST-REU-0003**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 4, 1996 14:20
PRIMSOURCE: SOURCE	
INCIDENT: DATE	March 4, 1996
INCIDENT: LOCATION	Tel Aviv
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: at least 13" "wounded: more than 100"
PERP: ORGANIZATION ID	"Hamas"

MESSAGE: ID	**TST-REU-0004**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 4, 1996 14:30
PRIMSOURCE: SOURCE	
INCIDENT: DATE	March 4, 1996
INCIDENT: LOCATION	Tel Aviv
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: at least 12" "wounded: 105"
PERP: ORGANIZATION ID	"Hamas"

Summarisation by fact extraction (Radev and McKeown 1998, CL)

- Reason over templates
- New templates are generated by combining other templates
- The most important template, as determined by heuristics, is chosen for generation
- Rules:
  - Change of perspective: If the same source reports conflicting information over time, report both pieces of information
  - Contradiction: If two or more sources report conflicting information, choose the one that is reported by independent sources
  - Addition: If additional information is reported in a subsequent article, include the additional information
  - Refinement: Prefer more specific information over more general one (name of a terrorist group rather than the fact that it is Palestinian)
  - Agreement: Agreement between two sources is reported as it will heighten the reader's confidence in the reported fact
  - Superset/Generalization: If the same event is reported from different sources and all of them have incomplete information, report the combination of these pieces of information
  - Trend: If two or more messages reflect similar patterns over time, these can be reported in one statement (e.g. three consecutive bombings at the same location)
  - No Information: Report the lack of information from a certain source when this would be expected
- Output summary, deep-generated:

> Reuters reported that 18 people were killed in a Jerusalem bombing Sunday. The next day, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel Radio. Reuters reported that the radical Muslim group Hamas had claimed responsibility for the act.

- Problem: domain-specificity built into the templates

A domain-inspecific method: text extraction

- Split text in units (paragraphs or sentences or text tiles)
- Assign each unit a score of importance/“extractworthiness”, using sentential and/or relational features
  - Sentential features of a unit can be calculated in isolation, e.g. number of TF/IDF words or location
  - Relational features of a unit are calculated in context of other units, e.g. unit with highest amount of shared terms
- Extract sentences with highest score verbatim as extract

External marking of “more important” material

- Text is globally structured (rhetorical sections, anecdotal/summary beginning in journalistic writing) – location feature
- Text is locally structured (paragraph structure; headlines and sub-headlines) – paragraph structure feature
- Important concepts/terms mark important prepositions – tf/idf feature
- Certain typographic regions are good places to find important concepts: captions, title, headlines – title feature
- Sentence length is important, but the experts argue; probably genre-dependent
- Phrases mark important sections (“in this paper”, “most important”) and less important sections (hedging by auxiliaries, adverbs) – cue phrase feature

Sentential features, I

1. Concept feature (Luhn, 1958)
   - Find concepts using *tf* (nowadays: *tf*idf*), sentence score = no of frequency concepts in sentence
2. Header feature (Baxendale, 1959)
   - Find concepts in title (variation: title and headlines), sentence score = no of title concepts in sentence
3. Location feature (Edmundson, 1969)
   - Divide text into n equal sections
   - sentences in section $1 \leq i \leq n$ get sentence score = $\frac{1}{i}$
   - Always used in combination

Sentential features, II

4. Paragraph feature
   - First sentence in paragraph gets a higher score than last one, and higher than sentences in the middle
   - Always used in combination
5. Cue phrases (Paice, 1991)
6. First-sentence-in-section feature
7. Sentence length
8. Occurrence of bonus or malus word (ADAM system, Zomora (1972))
9. Occurrence of a named entity (Kupiec et al., 1995)

Combination of sentential features: manually

- Combinations of features are more robust than single features
- Manual feature combination (Edmundson):

$$Score(S) = \alpha A + \beta B + ...\omega O$$

A, B,..O: feature scores
α, β, ω: manual weights

Combination of sentential features: machine learning

- Kupiec, Pedersen, Chen: A trainable document summariser, SIGIR 1995
- Create examples of sentences that are abstract-worthy, calculate their features, using 5 well-known features ($F_1 \ldots F_5$)
- Use Naive Bayesian classifier:

$$P(s \in S|F_1, \ldots, F_k) = \frac{P(F_1,...,F_k|s \in S)P(s \in S)}{P(F_1,...,F_k)} \approx \frac{P(s \in S)\prod_{j=1}^{k} P(F_j|s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

$P(s \in S\|F_1, \ldots, F_k)$:	Probability that sentence s from the source text is included in summary S, given its feature values;
$P(s \in S)$:	Probability that a sentence s in the source text is included in summary S unconditionally; compression rate of the task (constant);
$P(F_j\| \, s \in S)$:	probability of feature-value pair occurring in a sentence which is in the summary;
$P(F_j)$:	probability that the feature-value pair F_j (j th feature-value pair out of k feature-value pairs) occurs unconditionally;

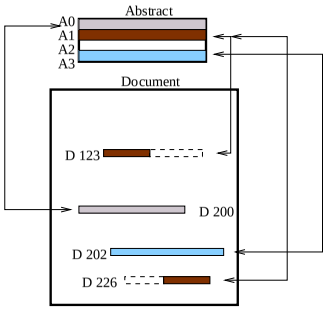
Finding the right gold standard

Subjective measures:

- Humans subjects select sentences (system developers?)

Looking for more objective measures:

- Earl: indexible sentences
- Kupiec et al: sentences with similarity to abstract sentences



Kupiec et al: gold standard

- Find best match for each abstract sentence by automatic similarity measure
- One example for a similarity measure is based on the longest common substring:

$$lcs(X,Y) = \frac{length(X) + length(Y) - edit_{i,d}(X,Y)}{2}$$

(where $edit_{i,d}$ is the minimum number of deletions and insertions needed to transform X into Y).

- Possible similarity measures are the ratio of longest common substring to the maximum length of the two sentences, or the average.
- Reject sentences with similary $<$ .5; accept sentences with similarity $>$ 0.8, hand-judge sentences with medium similarity $.5 \leq X \leq .8$

Kupiec et al's evaluation

- Corpus of 85 articles in 21 journals
- Extract as many sentences as there are gold standards in the document → precision = recall
- Very high compression makes this task harder
- Results:

Feature	Individual	Cumulative
Cue Phrases	33%	33%
Location	29%	42%
Sentence Length	24%	44%
*tf*idf*	20%	42%
Capitalization + *tf*idf*	20%	42%
Baseline	24%	

Example of an extract (Microsoft's AutoSummarize)

Distributional Clustering of English Sentences
Distributional Similarity To cluster nouns n according to their conditional verb distributions pn, we need a measure of similarity between distributions.
We will take (1) as our basic clustering model.
In particular, the model we use in our experiments has noun clusters with cluster memberships determined by p(njc) and centroid distributions determined by p(vjc).
Given any similarity measure d(n;c) between nouns and cluster centroids, the average cluster distortion is
If we maximize the cluster membership entropy
Clustering Examples
Figure 1 shows the five words most similar to the each [sic] cluster centroid for the four clusters resulting from the first two cluster splits.
Model Evaluation
1990. Statistical mechanics and phrase transitions in clustering.

Source: "*Distributional Clustering of English Sentences*" by Pereira, Tishby and Lee, ACL 1993

What are extracts good for?

- Extraction is the basis of all robust and reliable summarisation technology widely deployed nowadays
- It can give readers a rough idea of what this text is about
- Information analysts work successfully with them
- Task-based evaluation results:
  - Tombros et al. (1998) show slight improvement in precision and recall and larger improvement in time for a human search task
  - Mani et al. (1999) slight loss in accuracy and large advantage in time saving (50% of the time needed) for a relevance decision task

Problems with extracts

- Unclear to reader why particular sentence was chosen
- Coherence (syntactic, local problems)
  - Dangling anaphora
  - Unconnected discourse markers
- Cohesion (semantic discontinuities, global)
  - Concepts and agents are not introduced
  - Succession of events does not seem coherent

Fixes for coherence problems

- E.g. dangling anaphora:
  - resolve anaphora
  - recognize anaphoric use (as opposed to expletive use (“it”, Paice and Husk 1987)), then either
    - exclude sentences with dangling anaphora
    - include previous sentence if it contains the referent (Johnson et al. 1993; also for definite NPs) – But: length!
- There are no fixes for cohesion

Strategies for summary evaluation

1. Subjective judgements:
   How much do subjects like this summary? How coherent, well-written, etc do they find it?
2. Comparison to “gold standard” (predefined right answer):
   In how far does this summary resemble the “right answer”?
3. Task-based evaluation:
   How well can humans perform a task if they are given this summary?
4. Usability evaluation (extrinsic):
   Does the recipient of the summary have to change it? How much?

Problems in Summarisation Evaluation

1. Subjective judgements
   - Subjects can be biased
   - How to make sure they understand the same thing under ”informativeness”, for instance
2. Comparison to “gold standard”
   - by sentence co-selection, surface string similarity or “information overlap”
   - Problematic: humans do not agree on what a good summary is
   - Doubt about existence of a “gold standard”
3. Task-based evaluation
   - Probably the best evaluation around
   - Hard to define the task/set up the experiment
   - Time-consuming and expensive to do experiment
   - For final, end-of-project evaluation, not for day-to-day evaluation

Summary

- Summarisation by deep methods and problems
- Summarisation by text extraction
  - Importance features
  - Kupiec et al.’s (1995) method and training material
  - Lexical chains
- Summarisation evaluation and its problems