Information Extraction and Question-Answering Systems

Foundations and methods

Dr. Günter Neumann
LT-Lab, DFKI
neumann@dfki.de

22/02/2002 1

What the lecture will cover

- Machine Learning for IE
- Statistical Methods for lexical processing
- Evaluation Methods
- Basic Terms & Examples
- Parsing of Unrestricted Text
- Generic NL Core system
- Domain Modelling
- Question/Answering Core components
- Advanced Topics

22/02/2002 2

Evaluation

- How can we compare human and system performance?
- How can we measure and compare different methods?
- What can we learn for future system building?

Evaluation

- Information extraction
  - MUC: Message Understanding Conference
  - Languages considered:
    - English, Chinese, Spanish, Japanese
  - First round: 1987
- Textual Question/Answering
  - TREC: Text REtrieval Conference
  - Languages considered
    - English
  - First round: 1999

The Message Understanding Conference (MUC)

- Sponsored by the Defense Advanced Research Projects Agency (DARPA) 1991-1998.
- Developed methods for formal evaluation of IE systems
- In the form of a competition, where the participants compare their results with each other and against human annotators‘ key templates.
- Short system preparation time to stimulate portability to new extraction problems. Only 1 month to adapt the system to the new scenario before the formal run.

MUC: Evaluation procedure

- Corpus of training texts
- Specification of the IE task
- Specification of the form of the required output
- Keys: ground truth-human produced responses in output format
- Evaluation procedure
  - Blind test
  - System performance automatically scored against keys

MUC Tasks

- MUC-1 (87) and MUC-2 (89)
  - ➢ Messages about naval operations
- MUC-3 (91) and MUC-4 (92)
  - ➢ News articles about terrorist activity
- MUC-5 (93)
  - ➢ News articles about joint venture and microelectronics
- MUC-6 (95)
  - ➢ News articles about management changes
- MUC-7 (97)
  - ➢ News articles about space vehicle and missile launches

Events – Relations - Arguments

Examples of events or relationships to extract	Examples of their arguments
Terrorist attacks (MUC-3) (example corpus/output file)	Incident_Type, Date , Location, Perpetrator, Physical_Target, Human_Target, Effects, Instrument
Changes in corporate executive management personnel (MUC-6) (DFKI corpus German)	Post, Company, InPerson, OutPerson,VacancyReason,OldOrganisation, NewOrganisation
Space vehicles and missile launch events (rocket launches) (MUC-7)	Vehicle_Type, Vehicle_Owner, Vehicle_Manufacturer, Payload_Type, Payload_Func, Payload_Owner,Payload_Origin,Payload_Target, Launch, Date, Launch Site, Mission Type, Mission Function, etc.

Evaluation metrics

- Precision and recall:
  - Precision: correct answers/answers produced
  - Recall: correct answers/total possible answers
- F-measure
  - Where β is a parameter representing relative importance of P & R:
    $$F = \frac{(\beta^2 + 1)PR}{(\beta^2 P + R)}$$
  - E.g., β=1, then P&R equal weight, β=0, then only P
- Current State-of-Art: F=.60 barrier

22/02/2002 9

MUC extraction tasks

- Named Entity task (NE)
- Template Element task (TE)
- Template Relation task (TR)
- Scenario Template task (ST)
- Coreference task (CO)

22/02/2002 10

Named Entity task (NE)

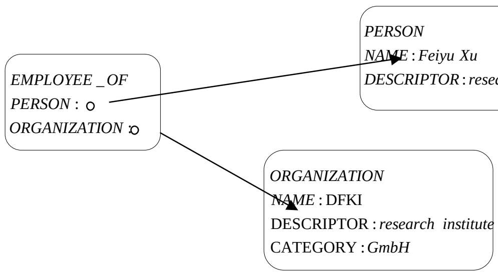
Mark into the text each string that represents, a person, organization, or location name, or a date or time, or a currency or percentage figure (this classification of NEs reflects the MUC-7 specific domain and task)

Template Element task (TE)

Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text (TE consists in generic objects and slots for a given scenario, but is unconcerned with relevance for this scenario)

Template Relation task (TR)

Extract relational information on employee_of, manufacture_of, location_of relations etc. (TR expresses domain-independent relationships between entities identified by TE)



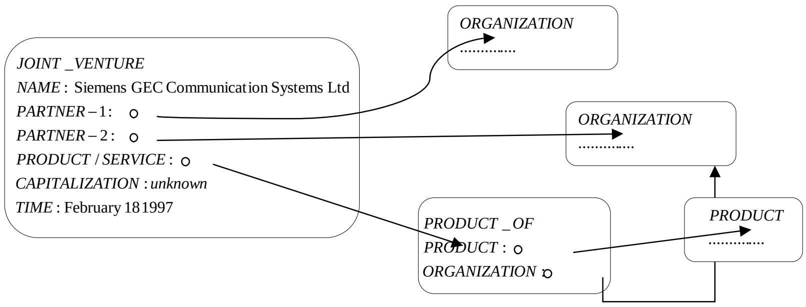
22/02/2002 13

Scenario Template task (ST)

Extract prespecified event information and relate the event information to particular organization, person, or artifact entities (ST identifies domain and task specific entities and relations)

22/02/2002 14

ST example



Coreference task (CO)

Capture information on corefering expressions, i.e. all mentions of a given entity, including those marked in NE and TE (Nouns, Noun phrases, Pronouns)

An Example

The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Head is a staff scientist at We Build Rockets Inc.

- NE: entities are *rocket*, *Tuesday*, *Dr. Head* and *We Build Rockets*
- CO: *it* refers to the rocket; *Dr. Head* and *Dr. Big Head* are the same
- TE: the rocket is *shiny red* and Head's *brainchild*
- TR: Dr. Head *works for* We Build Rockets Inc.
- ST: a *rocket launching event* occured with the various participants.

From: Tablan, Ursu, Cunningham, eurolan 2001

Scoring templates

- Templates are compared on a slot-by-slot basis
  - ➢Correct: response = key
  - ➢Partial: response ≈ key
  - ➢Incorrect: response ≠ key
  - ➢Spurious: key is blank
    - overgen=spurious/actual
  - ➢Missing: response is blank

Tasks evaluated in MUC 3-7

Eval\Task	NE	CO	RE	TR	ST
MUC-3					YES
MUC-4					YES
MUC-5					YES
MUC-6	YES	YES	YES		YES
MUC-7	YES	YES	YES	YES	YES

Maximum Results Reported in MUC-7

Meassure\Task	NE	CO	TE	TR	ST
Recall	92	56	86	67	42
Precision	95	69	87	86	65

Human on NE task	F	R	P
Annotator 1	98.6	98	98
Annotator 2	96.9	96	98

Human on ST task: ~ 80 % F

Details from MUC-7 online

TREC Question Answering Track

- Goal: motivate research on systems that retrieve answers rather than documents in response to a question
- Subject matter of questions is not restricted (open domain)
- Type of questions is limited to
  - Fact-based, short-answer questions
  - Answers are usually entities to information extraction systems (e.g., when, where, who, what, ...)
- So far, two QA TRECs have happend
  - TREC-8, November, 1999
  - TREC-9, November, 2000

Data used in QA-TREC-8/9

	TREC-8	TREC-9
# of dodcuments	528,000	979,000
MB of document text	1904	3033
Document sources	TREC disks 4-5: LA times, Financial times, FBIS, Federal Register	News from TREC disks 1-5: AP newswire, WSJ, San Jose Mercury News, Financial times, LA times, FBIS
# of questions released	200	693
# of questions evaluated	198	682
Question sources	FAQ finder log, assessors, participants	Encarta log, Excite log

TREC-8: Question source

- Most questions were from participants or NIST assessors
  - Main reason: FAQFinder logs not very usefull (rare relation to TREC document texts)
  - Questions often back-formulations of statements in the documents (made by participants!)
  - Questions therefore often unnatural
  - Easies QA task since target documents contained most of the questions words

TREC-9: Question source

- Only use query logs (no back-formulation)
- Encarta (MS): grammatical questions
- Excite log:
  - ➢Often ungrammatical
  - ➢But use words for formulating questions without reference to TREC documents

TREC-9 question variants

- Question variants
  - Syntactic paraphrases
  - Are QA system robust to the variety of different ways a question can be phrased?
- Problem: What counts as a real paraphrase?
  - What is Dick Clark‘s birthday? („November 29“) vs. When was Dick Clark‘s birthday („Nov. 29 + year“)
  - What is the location of the Orange Bowl? vs. What city is the Ornage Bowl in?

22/02/2002 25

The TREC QA Track: Task Definition

- Inputs:
  - ➢4GB newswire texts (from the TREC text collection)
  - ➢File of natural language questions, e.g.
    - *Where is the Taj Mahal?*
    - *How tall is the Eiffel Tower?*
    - *Who was Johnny Mathis' high school track coach?*

22/02/2002 26

The TREC QA Track: Task Definition

- Outputs:
  - Five ranked answers per question, including pointer to source document
    - 50 byte category
    - 250 byte category
  - Scoring function, e.g., Q/A word overlap count
  - Up to two runs per category per site
- Limitations:
  - Each question has an answer in the text collection
  - Each answer is a single literal string from a text (no implicit or multiple answers)

22/02/2002 27

The TREC QA Track: Metrics and Scoring

- The principal metric is **Mean Reciprocal Rank (MRR)**
  - ➢Correct answer at rank 1 scores 1
  - ➢Correct answer at rank 2 scores 1/2
  - ➢Correct answer at rank 3 scores 1/3
  - ➢…
  - Sum over all questions and divide by number of questions

22/02/2002 28

The TREC QA Track: Metrics and Scoring

- More formally: $MRR = \frac{\sum_{i=1}^{N} r_i}{N}$

  where N = # questions, r_i = the reciprocal of the best (lowest) rank assigned by a system at which a correct answer is found for question i, or 0 if no correct answer was found

- Judgements made by human judges based on answer string alone (lenient evaluation) and by reference to documents (strict evaluation)

TREC-9 QA track result

Participants: 20
Short answer types: MRR between 0.58 - 0.10

Participant	MRR	# not found
Southern Methodist U.	0.58	229 (34%)
ISI, U. of S. Cal.	0.32	385 (57%)
MultiText, U. Waterloo	0.32	395 (58%)
…	…	…
LIMSI	0.18	499 (73%)
CL Research	0.14	550 (81%)
Seoul National U	0.10	577 (85%)

TREC-9 QA track result

Participants: 20
Long answer types: MRR between 0.76 – 0.30

Participant	MRR	# not found
Southern Methodist U.	0.76	95 (14%)
IBM (Ittycheriah)	0.46	263 (39%)
Queens College, CUNY	0.46	264 (39%)
...	...	...
KAIST	0.33	362 (53%)
National Taiwan U	0.32	376 (55%)
CL Research	0.30	386 (57%)

Automatic evaluation is still a problem

- Different QA runs seldom return exactly the same answer string
- Difficult: difference of a new string and a judged string is difficult to determine automatically (note, an automatic solution would require a system which is able to prove that two different strings „mean“ the same answer)
- Approximate solution:
  - from a set of judged answers create an question pattern.
- Then any answer string that matches any pattern for its question is marked correct.

Example of question pattern (as Perl expressions)

Who was Jane Goodall?
Naturalist
Chimpanzee\s+specialist
Chimpanzee\s*-?\s*observer
Pioneered.*study\s+of\s+primates
Ethnologist
Animal\s+behaviorist
...

\s = whitespace character

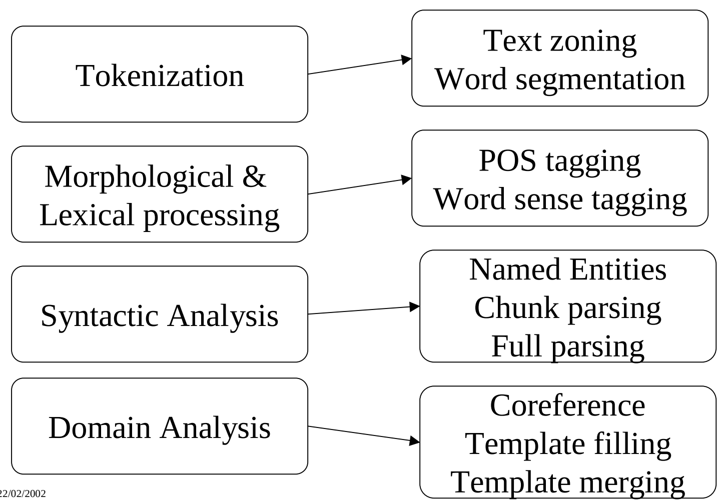
Multiple-answer occurences

- A document contains several plausible answers for a question
  - What does Peugeot company manufacture?
    - Trucks, cars, motors
- Problem:
  - Three individual answers
  - One complex answer
  - How to find out semantic relationships?

The Potential of NLP for Question Answering

- NLP has failed to deliver significant improvements in the document retrieval task.
  - ➢ Will the same be true of QA?
- Must depend on the definition of task
  - ➢ Current TREC QA task is best construed as micro passage retrieval
- There are a number of linguistic phenomena relevant to QA which suggest that NLP ought to be able to help, in principle.
- But, it also now seems clear from TREC-9 results that NLP techniques do improve the effectiveness of QA systems in practice.

A Bare-Bones Text Extraction System



Text zoning

- Parse text into segments
  - Separate formatted from unformatted text regions
    - email subject, body
    - Title, sections, paragraphs, sentence, HTML-tables
  - Problem: semi-formal ascii texts, e.g., talk announcements
- Advanced systems (Choi et.al, EMNLP-2001)
  - Identify elementary blocks (smallest text segments that can describe an entire topic, e.g., sentences, paragraphs, ...)
  - Similarity metric estimates the likelihood of two segments describing the same topic (based on Latent Semantic Analysis)
- Usefull for text zooming:
  - Answer extraction (paragraph indexing)
  - Coreference tasks (coreference chains)
  - Text mining (topic maps)

Word segmentation

- Tokenization: isolation of word-like units from text
- Results in two types of tokens
  - Units, whose character structure is recognizable (e.g., punctuation, numbers, date, ...)
  - Units, which will go morpholical analysis
- Computationally simple: regular grammars
- So, is it a problem? Yes, say Grefenstette & Tapanainen, Complex'94

Problems of tokenization

- Isolation of word and sentence boundaries involves the use of ambiguous punctuation
- Major problem: the dot
  e.g., Brown Corpus
  - 52511 sentences ended by a full stop (. or ?)
  - 3569 contain at least one non-terminal period
  - 93.20% accuracy, if every dot is interpreted as full stop
- Can be improved by adding increasing levels of linguistic sophistication

By way: often, very simple (word-based) methods yield about 90% correctness;
the challenge are mostly the remaining 10%; the real challenge are then the last 5%.

Ambiguous Separators in Numbers

- Ambiguous comma & period
  - English: 123,456.78; RE: ([0-9])+[,])*[0-9]([.][0-9]+)?
  - French: 123 456,78; RE: ([0-9])+[ ])*[0-9]([,][0-9]+)
- Some other english expressions
  - [0-9]+(\/[0-9]+)+ Fractions, Dates
  - ([+\-])?[0-9]+(\.)?[0-9]*% Percent
  - ([0-9]+,?)+(\.[0-9]+|[0-9]+)* Decimal Numbers
- Improves correction:
  - Only 3340 recognized incorrectly
  - From 93.30% to 93.64%

Abbreviations & lexicon

- Heuristic: any period not followed by blank is not a full stop
  - Yields 93.78%
- Analyze structure of abrev.
  - A., B., U.S., m.p.h.
  - Mr., St.
  - Yields: 97.66%
- Using lexicon & morphology: 98.27%
- Palmer & Hearst (1994):
  - Neural net applied to morphologically tagged text
  - 98.5 % success rate (not making use of capitalization)
- Other problems
  - Feld, Wiesen-, und Stallhasen
  - Mixed epressions: 12:30 h vs. 12:30 Uhr
  - Noise: mph vs. m.p.h.

Lexical data base

- Words of a language together with morpho-syntactic, syntactic and semantic information
- In general, the lexial attributes describe all possible readings
- Usually contains non-compositional, lexicalized expressions
- Size of the lexical units determined also by computational processes available
- Usually, lexicon elements are normalized entries which characterize a set of common word forms (e.g., „haus“ for „Häuser“, „Häusern“, „Hauses“, ...)
- Normalized units also called *lemmas*

POS tagging & morphological processing

- Goal: map word form to lexical entries & externalize implicitly available information
- Tasks:
  - ➢ Find part-of-speech
  - ➢ Analyse inflection
  - ➢ Compound/derivation analysis
- Problem: Mehrdeutigkeit
  - ➢ *Ich meine meine Tasche*
  - ➢ *Stau-becken* vs. *Staub-ecken*
  - ➢ *Bank* (Sitzgelegenheit vs. Geldinstitut)
- Computational feasible
  - ➢ Finite state technology
  - ➢ Statistical-based disambiguation methods

Example of inflection analysis

(based on Morphix feature output)

String	**Stamm**	**POS**	**Gender/ Person**	**Fall**	**Nummer**	**Form**
Nach	Nach	Prep		Dat		
dem	d	Det	m	Dat	Sg	
			n	Dat	Sg	
Kauf	kauf	Noun	m	Nom	Sg	
			m	Dat	Sg	
			m	Acc	Sg	
		Verb			Sg	Imperativ
weiterer	weit	Adj	m	Nom	Sg	
			m	Gen	Pl	
			f	Gen	Sg	
			f	Gen	Pl	
Anteile	anteil	Noun	m	Nom	Pl	
			m	Gen	Pl	
			m	Acc	Pl	
			m	Dat	Sg	
halten	halt	Noun	m	Dat	Pl	
		Verb	Anrede		Sg	
		Verb	Anrede		Pl	
		Verb	1.P		Pl	
		Verb	3. P		Pl	
wir	wir	PersPron		Nom	Pl	

Syntactic Analysis is heavy!

- It is assumed that NL-syntax has more than context free power
- Problems
  - Free word order
    - *Peter sieht den Mann vs. Den Mann sieht Peter*
  - Discontinous elements
    - *Ich sage das Treffen ab.*
  - Elliptical and anaphoral expressions
    - *Er sah den Man, wie er den schweren Weg hinauf kam.*
    - *Drei europäische Sprachen werden von zehn Linguisten gesprochen, zwei asiatische auch.*

Highly ambiguous utterances

Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften. (Uszkoreit)

- Lexical and morphological ambiguity (32)
- Morphosyntactic ambiguity (case ambiguity) (8)
- Attachment ambiguity (252)
  - PP attachment (63)
  - Extraposed relative clauses (4)

=64.512 readings

Parsing of free texts

- Only parts of a text are of interest
- Real sentence can be really long (>100 words)
- Parts of the syntactic structure might be expressed via text structure (items)
- Creative use of language (mixed style/ languages)
- Mass of technical terms
- Ungrammatical/telegram-like style
- Syntactic analysis must be fast

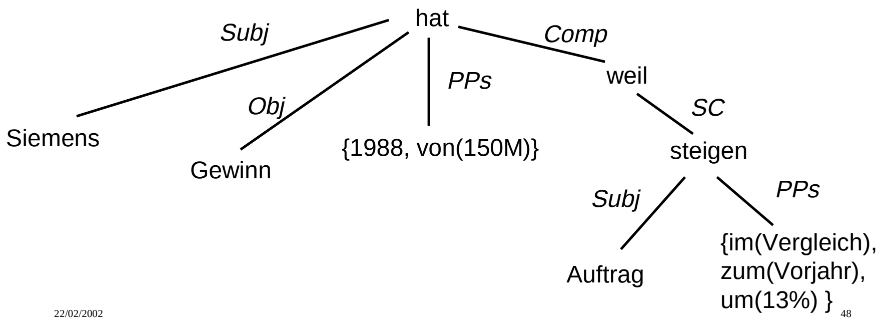
- How to apply syntactic analyis only on interesting parts of a sentence?
- How to obtain near-deterministic speed?
- How to obtain robustness?
- Is there any way to obtain system adaptation?

- Identify modules according to type and complexity of syntactic units
  - Fine-grained precision decisions
- Try to apply FST where possible
- Use corpus-based mechanism

Cascaded Chunk parsing

- Recognition of
  - Named entities
  - General phrases (nominal prepositional, verb phrases)
  - Grammatical function

[$_{PN}$Die Siemens GmbH] [$_{V}$hat] [$_{year}$1988][$_{NP}$einen Gewinn] [$_{PP}$von 150 Millionen DM], [$_{Comp}$weil] [$_{NP}$die Auftraege] [$_{PP}$im Vergleich] [$_{PP}$zum Vorjahr] [$_{Card}$um 13%] [$_{V}$gestiegen sind].



Coreference resolution

- Goal: find different verbalizations of the same entity;
  needed for template merging

- Example:

  Da flüchten *sich die einen* ins Ausland, wie etwa *der Münchner Strickwarenhersteller März GmbH* oder *der badische Strumpffabrikant Arlington Socks, GmbH*. Ab kommendem Jahr strickt *März* knapp drei Viertel seiner Produktion in Ungarn.

  (*Therefore some take refuge abroad, like the Münchner knitware producer März GmbH or the badische Strumpffabrikant Arlington Socks, GmbH. From next year on, März knits around three quarters of its production in Hungary.)*

- Modular approach needed
  - handle nominal reference problems with actual available structural information as early as possible on different processing levels

22/02/2002 49

Complexity Factors

- Language
  - Orthography
  - Morphology
- Genre
  - Case
  - Formality
    - Newspapers
    - Email
    - speech
- Text
  - Length
  - Non-textual data
    - Tabular data
    - Graphical data
- Task
  - MUC tasks

22/02/2002 50

IE: compromise NLP

- Task characteristic
  - Lots of texts
  - Dirty texts
  - World knowledge needed
- Compromise
  - Finite-state models
  - Robust techniques
  - Domain specific processing at each stage of analysis

The bottom line:
Find the most favorable tradeoff between recall and precision for the task at hand.