Information Extraction

Katharina Kaiser
http://www.ifs.tuwien.ac.at/~kaiser

Information Extraction

- Definition
- History
- Architecture of IE systems
- Wrapper systems
- Approaches
- Evaluation

IE: Definition

- Natural Language Processing (NLP)
  1. Process unstructured, natural language text
  2. Locate specific pieces of information in the text
  3. Fill a database
- Wrapper technology
  1. [Retrieve information from different repositories]
  2. Merge and unify them
  3. Unify them

IE: History

- Message Understanding Conferences (MUC)

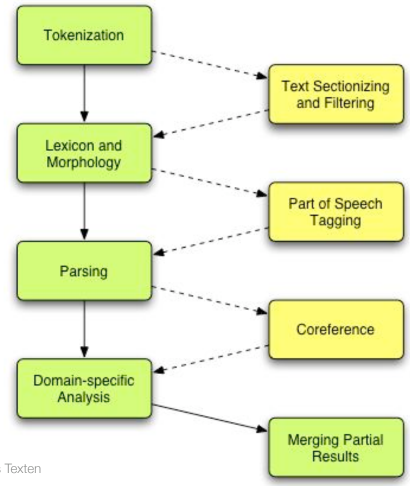
Conference	Year	Text	Domain
MUC-1	1987	military communications	naval operations messages
MUC-2	1989	military communications	naval operations messages
MUC-3	1991	news	terrorism in Latin American Countries
MUC-4	1992	news	terrorism in Latin American Countries
MUC-5	1993	news	joint ventures and microelectronics domain
MUC-6	1995	news	management changes
MUC-7	1997	news	satellite launch reports, air crash,

- Text REtrieval Conferences (TREC)

Information Extraction

- Definition
- History
- Architecture of IE systems
- Wrapper systems
- Approaches
- Evaluation

IE: Architecture



IE: Architecture

- Tokenization module

IE: Architecture

- Lexicon and Morphology module

<ENAMEX TYPE=„LOCATION“>Italy</ENAMEX>‘s business world was rocked by
the announcement <TIMEX TYPE=„DATE“>last Thursday</TIMEX> that Mr.
<ENAMEX TYPE=„PERSON“>Verdi</ENAMEX> would leave his job as vice-president
of <ENAMEX TYPE=„ORGANIZATION“>Music Masters of Milan, Inc</ENAMEX>
to become operations director of
<ENAMEX TYPE=„ORGANIZATION“>Arthur Andersen</ENAMEX>.

IE: Architecture

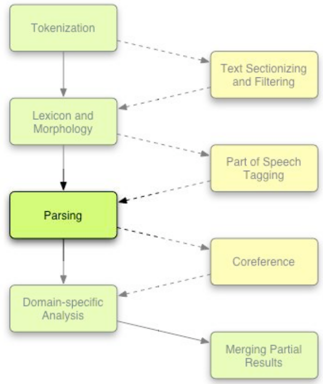
- Part of Speech (POS) Tagging

  1. Use lexicon containing words and possible POS tags
  2. Guess POS tag of unknown words
  3. Words with multiple/dubious tags: seeking of most likely tag

IE: Architecture

- Parsing
  - Syntactic Analysis

[Bridgestone Sports Co.]NG [said]VG
[Friday]NG [it]NG [has set up]VG
[a joint venture]VG [in]P [Taiwan]NG
[with]P [a local concern]NG [and]P
[a Japanese trading house]NG
[to produce]VG [golf clubs]NG
[to be shipped]VG [to]P [Japan]NG.



IE: Architecture

- Coreference

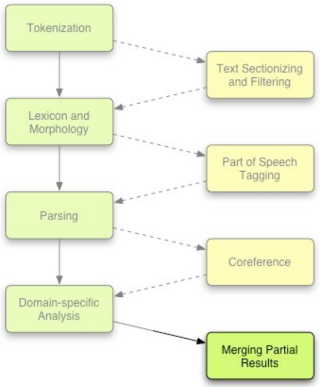
[Bridgestone Sports Co.]NG [said]VG
[Friday]NG [it]NG [has set up]VG
[a joint venture]VG [in]P [Taiwan]NG
[with]P [a local concern]NG [and]P
[a Japanese trading house]NG
[to produce]VG [golf clubs]NG
[to be shipped]VG [to]P [Japan]NG.

IE: Architecture

- Domain-specific Analysis
  - Templates: consist of a collection of slots (i.e., attributes)
  - Values: original text, one or more of a finite set of predefined alternatives, or pointers to other templates slots
  - Domain specific extraction patterns:
    1. atomic approach
    2. molecular approach

IE: Architecture

- Merging Partial Results



Information Extraction

- Definition
- History
- Architecture of IE systems
- Wrapper systems
- Approaches
- Evaluation

IE: Wrapper Systems

- Different structure of each document, sites change periodically
- “Wrapper generation”, “Wrapper maintenance”
- Format uniqueness and completeness
- HTML-quality level

IE: Wrapper Systems

- Format uniqueness and completeness
  - Rigorous structure: unique format and complete information
  - Semi-rigorous structure: unique format and incomplete information
  - Semi-relaxed structure: no unique format and complete information
  - Relaxed structure: no unique format and incomplete information

IE: Wrapper Systems

- HTML-quality level
  - High level: each item in the result page is surrounded by a couple of HTML tags, such as -; each tagged item corresponds to exactly one attribute of the original data
  - Low level: a string between two HTML tags corresponds to more than one output attribute; additional plain-text separators like ".", ",", ";" are used for separating the different attributes.
    An analysis of the HTML structure is not enough: a plain text analysis must be done

Information Extraction

- Definition
- History
- Architecture of IE systems
- Wrapper systems
- Approaches
- Evaluation

IE: Approaches

- Knowledge Engineering
- Automatic Learning
  - Supervised
  - Semi-supervised
  - Unsupervised

IE: Approaches

- Developing extraction rules
  - Learning rules using set of training examples
  - Reaching a state where we will able to extract correct information from other examples
  - Compromise: bias - variance
    - Bias: model does not follow the right trend in the data (*underfitting*)
    - Variance: model fits the data too closely (*overfitting*)

IE: Approaches

- Recall/Precision
  - POS: total possible correct responses
  - COR: number of correct values
  - INC: number or incorrect values
  - OVG: overgenerated values
- Statistical measures
- Interdependent pair
- Trade-off: only one measure can be optimized at the cost of the other

$$Recall = \frac{COR}{POS}$$

$$Precision = \frac{COR}{COR + INC + OVG}$$

IE: Knowledge Engineering

- FASTUS [early 1990's]
  - Finite State Automaton Text Understanding System

  1. Triggering
  2. Recognizing phrases
  3. Recognizing patterns
  4. Merging of incidents

  - Large dictionary

IE: Knowledge Engineering

- GE NLTOOLSET [1992]
  - Knowledge-based, domain-independent

  1. Pre-processing
  2. Linguistic analysis
  3. Post-processing

  - Lexicon: > 10,000 entries
  - Core grammar: 170 rules

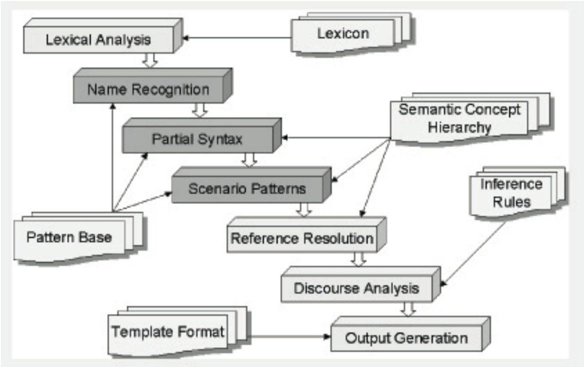
IE: Knowledge Engineering

- PLUM (Probabilistic Language Understanding Model) [1992]

  1. Pre-processing
  2. Morphological analysis
  3. Parsing
  4. Semantic interpreter
  5. Discourse processing
  6. Template generation

IE: Knowledge Engineering

- PROTEUS [1998]



IE: Automatic Learning

- Supervised learning systems
  - Input: set of (annotated) documents
  - Output: set of extraction patterns
  - Methods: Machine Learning (ML) techniques
  - Almost no knowledge about domain

IE: Automatic Learning

Linguistic Pattern
<subject> passive-verb
<subject> active-verb
<subject> verb infinitive
<subject> auxiliary noun

passive-verb <direct-object>
active-verb <direct-object>
infinitive <direct-object>
...

- AutoSlog [Riloff, 1993]
  - Extracts a domain-specific dictionary of concept nodes
    - Concept node: rule including a “trigger” word or word and a semantic constraint
  - Trigger in text and concept node’s condition satisfied: activate concept node and extract concept node definition
  - Single-slot extraction; no merging of similar concept nodes; only free text

IE: Automatic Learning

- PALKA [Kim and Moldovan, 1995]
  - Extraction rule: pair of meaning frame and phrasal pattern (Frame-Phrasal pattern structure (FP-structure))
  - New positive instance: new rule is generalized with existing ones
  - Avoid negative instance: existing rules are specialized

IE: Automatic Learning

- WHISK [Soderland, 1999]
  - Covering algorithms [Michalski, 1983]
  - Regular expressions in top-down induction
    1. Begins with most general rules
    2. Progressively specializing available rules
    3. Until set of rules cover all positive training examples
  - Post-pruning

IE: Automatic Learning

- RAPIER (Robust Automated Production of Information Extraction Rules) [Califf & Mooney, 1999]
  - Input: sample documents, filled templates
  - Output: pattern-match rules
  - Bottom-up learning algorithm (prefer high precision by preferring more specific rules)
  - Single-slot extraction; semi-structured text

IE: Automatic Learning

- GATE [Cunningham et al.,2002]
  - ANNIE (A Nearly New IE system)
    - Tokenizer
    - Sentence splitter
    - POS tagger
    - Gazetteer
    - Finite state transducer
    - Orthomatcher
    - Coreferencer

IE: Automatic Learning

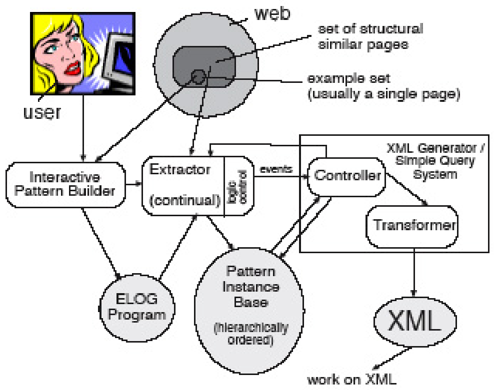
- WIEN (Wrapper Induction ENvironment) [Kushmerick et al., 1997]
  - Influenced by ShopBot [Doorenbos et al. 1997]
  - Bottom-up induction algorithm
  - Input: set of labeled pages
  - Delimiters must immediately precede and follow the data to be extracted
  - Cannot handle sources with missing items or items in varying order

IE: Automatic Learning

- Lixto [Baumgartner et al., 2001]
  - Declarative wrapper program for supervised wrapper generation
  - Visual and interactive user interface
  - Flexible hierarchical extraction pattern definition
  - Various kinds of conditions (e.g., contextual, internal, range conditions, references)
  - Internal rules language ELOG
  - No working inside the HTML source or tree representation

IE: Automatic Learning

- Lixto [Baumgartner et al., 2001]



IE: Automatic Learning

- Semi-supervised learning
  - Bootstrapping methods: expanding an initial small set of extraction patterns
- Unsupervised learning
  - Statement of the required information

IE: Automatic Learning

- Mutual Bootstrapping [Riloff & Jones, 1999]
  - Co-training algorithm using mutual bootstrapping for lexical discovery
  - Assumption
    a. Good pattern can find a good lexicon
    b. Good Lexicon can find good pattern

1. Initial data: a handful of lexical data
2. Patterns are discovered by the initial lexicon
3. Patterns are ranked and most reliable are used to extract more lexical items

IE: Automatic Learning

- EXDISCO [Yangarber et al., 2000]
  - Assumption
    a. Presence of relevant documents indicates good patterns
    b. Good patterns can find relevant documents
  1. Start: Unannotated corpus and handful of seed patterns
  2. Divide document set in “relevant document set” and “non-relevant document set”
  3. Generate “candidate patterns” from clauses in documents and rank patterns in correlation with relevant documents
  4. Add highest pattern to pattern set and re-rank each document using newly obtained pattern set

IE: Automatic Learning

- QDIE [Sudo, 2004]
  - Input: set of keywords
  - Parsing document by dependency parser and Named Entity tagger
  - Retrieves relevant documents specified by user’s query
  - Dependency trees of sentences: pattern extraction

IE: Automatic Learning

- RoadRunner [Crescenzi et al., 2001]
  - Automatically extract data from Web sources by exploiting similarities in page structure across multiple pages
  - Inducing the grammar of Web pages by comparing several pages containing long lists of data
  - Works well on data-intensive sites

IE: Choosing the Approach

- Knowledge Engineering vs. Automatic Learning
  - Availability of training data
  - Availability of linguistic resources
  - Availability of knowledge engineers
  - Stability of the final specifications
  - Level of performance required

Information Extraction

- Definition
- History
- Architecture of IE systems
- Wrapper systems
- Approaches
- Evaluation

IE: Evaluation

- Recall/Precision
  - POS: total possible correct responses
  - COR: number of correct values
  - INC: number or incorrect values
  - OVG: overgenerated values
- F-measure: geometric means
  - P ... precision
  - R ... recall
  - β ... weight parameter

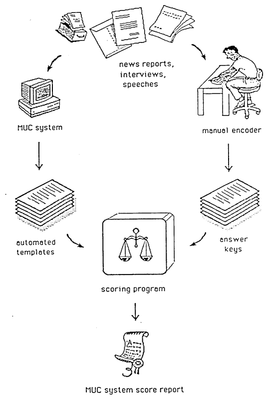
$$Recall = \frac{COR}{POS}$$

$$Precision = \frac{COR}{COR + INC + OVG}$$

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

IE: Evaluation

- Procedure



IE: Evaluation

TST1-MUC3-0004

BOGOTA, 30 AUG 89 (INRAVISION TELEVISION CADENA 2) -- [TEXT] LAST NIGHT'S TERRORIST TARGET WAS THE ANTIOQUIA LIQUEUR PLANT. FOUR POWERFUL ROCKETS WERE GOING TO EXPLODE VERY CLOSE TO THE TANKS WHERE 300,000 GALLONS OF THE SO-CALLED CASTILLE CRUDE, USED TO OPERATE THE BOILERS, IS STORED. THE WATCHMEN ON DUTY REPORTED THAT AT 2030 THEY SAW A MAN AND A WOMAN LEAVING A SMALL SUITCASE NEAR THE FENCE THAT SURROUNDS THE PLANT. THE WATCHMEN EXCHANGED FIRE WITH THE TERRORISTS WHO FLED LEAVING BEHIND THE EXPLOSIVE MATERIAL THAT ALSO INCLUDED DYNAMITE AND GRENADE ROCKET LAUNCHERS, METROPOLITAN POLICE PERSONNEL SPECIALIZING IN EXPLOSIVES, DEFUSED THE ROCKETS. SOME 100 PEOPLE WERE WORKING INSIDE THE PLANT.

THE DAMAGE THE ROCKETS WOULD HAVE CAUSED HAD THEY BEEN ACTIVATED CANNOT BE ESTIMATED BECAUSE THE CARIBE SODA FACTORY AND THE GUAYABAL RESIDENTIAL AREA WOULD HAVE ALSO BEEN AFFECTED.

THE ANTIOQUIA LIQUEUR PLANT HAS RECEIVED THREATS IN THE PAST AND MAXIMUM SECURITY HAS ALWAYS BEEN PRACTICED IN THE AREA. SECURITY WAS STEPPED UP LAST NIGHT AFTER THE INCIDENT. THE LIQUEUR INDUSTRY IS THE LARGEST FOREIGN EXCHANGE PRODUCER FOR THE DEPARTMENT.

TST1-MUC3-0004

BOGOTA, 30 AUG 89 (INRAVISION TELEVISION CADENA 2) -- [TEXT] LAST NIGHT'S TERRORIST TARGET WAS THE ANTIOQUIA LIQUEUR PLANT. FOUR POWERFUL ROCKETS WERE GOING TO EXPLODE VERY CLOSE TO THE TANKS WHERE 300,000 GALLONS OF THE SOCALLED CASTILLE CRUDE, USED TO OPERATE THE BOILERS, IS STORED. THE WATCHMEN ON DUTY REPORTED THAT AT 2030 THEY SAW A MAN AND A WOMAN LEAVING A SMALL SUITCASE NEAR THE FENCE THAT SURROUNDS THE PLANT. THE WATCHMEN EXCHANGED FIRE WITH THE TERRORISTS WHO FLED LEAVING BEHIND THE EXPLOSIVE MATERIAL THAT ALSO INCLUDED DYNAMITE AND GRENADE ROCKET LAUNCHERS, METROPOLITAN POLICE PERSONNEL SPECIALIZING IN EXPLOSIVES, DEFUSED THE ROCKETS. SOME 100 PEOPLE WERE WORKING INSIDE THE PLANT.

THE DAMAGE THE ROCKETS WOULD HAVE CAUSED HAD THEY BEEN ACTIVATED CANNOT BE ESTIMATED BECAUSE THE CARIBE SODA FACTORY AND THE GUAYABAL RESIDENTIAL AREA WOULD HAVE ALSO BEEN AFFECTED.

THE ANTIOQUIA LIQUEUR PLANT HAS RECEIVED THREATS IN THE PAST AND MAXIMUM SECURITY HAS ALWAYS BEEN PRACTICED IN THE AREA. SECURITY WAS STEPPED UP LAST NIGHT AFTER THE INCIDENT. THE LIQUEUR INDUSTRY IS THE LARGEST FOREIGN EXCHANGE PRODUCER FOR THE DEPARTMENT.

0. MESSAGE ID	TST1-MUC3-0004
1. TEMPLATE ID	1
2. DATE OF INCIDENT	29 AUG 89
3. TYPE OF INCIDENT	ATTEMPTED BOMBING
4. CATEGORY OF INCIDENT	TERRORIST ACT
5. PERPETRATOR: ID OF INDIV(S)	"MAN" "WOMAN"
6. PERPETRATOR: ID OF ORG(S)	-
7. PERPETRATOR: CONFIDENCE	-
8. PHYSICAL TARGET: ID(S)	"ANTIOQUIA LIQUEUR PLANT" "LIQUEUR PLANT"
9. PHYSICAL TARGET: TOTAL NUM	1
10. PHYSICAL TARGET: TYPE(S)	COMMERCIAL: "ANTIOQUIA LIQUEUR PLANT" "LIQUEUR PLANT"
11. HUMAN TARGET: ID(S)	"PEOPLE"
12. HUMAN TARGET: TOTAL NUM	PLURAL
13. HUMAN TARGET: TYPE(S)	CIVILIAN: "PEOPLE"
14. TARGET: FOREIGN NATION(S)	-
15. INSTRUMENT: TYPE(S)	*
16. LOCATION OF INCIDENT	COLOMBIA: ANTIOQUIA (DEPARTMENT)
17. EFFECT ON PHYSICAL TARGET(S)	NO DAMAGE: "ANTIOQUIA LIQUEUR PLANT" "LIQUEUR PLANT"
18. EFFECT ON HUMAN TARGET(S)	NO INJURY OR DEATH: "PEOPLE"

IE: Evaluation

SLOT	POS	ACT	COR	PAR	INC	ICR	IPA	SPU	MIS	NON	REC	PRE	OVG	FAL
template-id	113	215	107	0	0	0	0	108	6	17	95	50	50	
incident-date	109	103	56	21	26	0	21	0	6	4	61	64	0	
incident-type	113	107	77	20	10	0	20	0	6	0	77	81	0	0
category	81	67	55	0	8	0	0	4	18	28	68	82	6	8
indiv-perps	95	54	27	4	10	3	4	13	54	43	30	54	24	
org-perps	68	51	35	0	6	0	0	10	27	45	51	69	20	
perp-confidence	68	51	20	3	18	0	3	10	27	45	32	42	20	4
phys-target-ids	54	30	14	3	5	4	3	8	32	74	29	52	27	
phys-target-num	37	20	13	0	6	0	0	1	18	75	35	65	5	
phys-target-types	54	30	15	3	4	5	3	8	32	74	30	55	27	1
human-target-ids	144	95	50	14	17	4	14	14	63	16	40	60	15	
human-target-num	92	76	45	1	25	0	1	5	21	16	49	60	6	
human-target-types	144	95	54	21	6	2	21	14	63	16	45	68	15	1
target-nationality	18	6	4	1	0	3	1	1	13	99	25	75	17	0
instrument-types	25	11	6	0	1	0	0	4	18	84	24	54	36	0
incident-location	113	107	56	37	14	0	0	0	6	0	66	70	0	
phys-effects	36	18	12	2	2	3	2	2	20	89	36	72	11	0
human-effects	55	34	14	7	2	3	7	11	32	72	32	51	32	1
MATCHED ONLY	1361	1170	660	137	160	27	100	213	404	751	54	62	18	
MATCHED/MISSING	1419	1170	660	137	160	27	100	213	462	797	51	62	18	
ALL TEMPLATES	1419	1929	660	137	160	27	100	972	462	1926	51	38	50	
SET FILLS ONLY	594	419	257	57	51	16	57	54	229	507	48	68	13	0

Scoring Key:
POS (POSSIBLE) - the number of slot fillers according to the key target templates
ACT (ACTUAL) - the number of slot fillers generated by the system (= COR + PAR + INC + SPU)
COR (CORRECT) - the number of correct slot fillers generated by the system
PAR (PARTIAL) - the number of partially correct slot fillers generated by the system
INC (INCORRECT) - the number of incorrect slot fillers generated by the system
ICR (INTERACTIVE CORRECT) - the subset of COR judged correct during interactive scoring
IPA (INTERACTIVE PARTIAL) - the subset of PAR judged partially correct during interactive scoring
SPU (SPURIOUS) - the number of spurious slot fillers generated by the system
MIS (MISSING) - the number slot fillers erroneously not generated by the system
NON (NONCOMMITTAL) - the number of slots that were correctly left unfilled by the system
REC (RECALL) - the ratio of COR plus (.5 x) PAR slot fillers to POS slot fillers
PRE (PRECISION) - the ratio of COR plus (.5 x) PAR slot fillers to ACT slot fillers
OVG (OVERGENERATION) - the ratio of SPU slot fillers to ACT slot fillers
FAL (FALLOUT) - the ratio of INC plus SPU slot fillers to the number of possible incorrect slot fillers (a complex formula)

IE: Evaluation

	Scenario Template Task	Named Entity Task	Template Element Task	Coreference Task	Template Relation Task
MUC-3	R < 52 % P < 58 % F < 46 %				
MUC-4	R < 59 % P < 59 % F < 56 %				
MUC-5	R < 59 % P < 60 % F < 52 %				
MUC-6	R < 59 % P < 72 % F < 57 %	R < 96 % P < 97 % F < 97 %	R < 77 % P < 88 % F < 80 %	R < 63 % P < 72 % F < 65 %	
MUC-7	R < 50 % P < 59 % F > 51 %	R < 92 % P < 95 % F < 94 %	R < 87 % P < 87 % F < 87 %	R < 79 % P < 59 % F < 62 %	R < 67 % P < 87 % F < 76 %
HUMAN F-Score (MUC-7)	85.15 % - 96.64 %	96.95 % - 97.60 %			

IE: Application Areas

- Database population
- Ontology population/evolving
- Text summarization
- ...