Ontology Learning (from text!)

Marie-Laure Reinberger
marielaure.reinberger@ua.ac.be
CNTS

Outline

- Definitions and description
- Machine Learning and Natural Language Processing for Ontology Learning
- Ontology Building Applications

Part I
Definitions and description

What's (an) ontology?

- Branch of philosophy which studies the nature and the organization of reality
- Structure that represents a domain knowledge (the meaning of the terms and the relations between them) to provide to a community of users a common vocabulary on which they would agree

What about: Thesauri – Semantic lexicons – Semantic networks ?

- Thesauri: standard set of relations between words or terms
- Semantic lexicons: lexical semantic relations between words or more complex lexical items
- Semantic networks: broader set of relations between objects

➢ Differ in the type of objects and relations

Thesaurus: example

- Roget: thesaurus of English words and phrases
  - groups words in synonym categories or concepts
- Sample categorization for the concept "Feeling":

AFFECTIONS IN GENERAL
  Affections
    Feeling
      warmth, glow, unction, vehemence;
      fervor, fervency;
      heartiness, cordiality;
      earnestness, eagerness;
      empressment, gush, ardor, zeal, passion...

Thesaurus: example

- MeSH (Medical Subject Headings)
  - provides for each term term variants that refer to the same concept
- MH= gene library

bank, gene	banks, gene
DNA libraries	gene banks
gene libraries	libraries, DNA
libraries, gene	library, DNA
library, gene	

Semantic lexicon: example

- WordNet: set of semantic classes (synsets)
- {board, plank}, {board, committee}
- tree
  woody_plant ligneous_plant
  vascular_plant tracheophyte
  plant flora plant_life
  life_form organism_being living_thing
  entity something
- tree tree_diagram
  … abstraction

Semantic network: example

- UMLS: Unified Medical Language System
- Metathesaurus: groups term variants that correspond to the same concept

HIV
  HTLV-III
  Human Immunodeficiency Virus
  ...

Semantic Network: example

- UMLS: Unified Medical Language System
- Semantic Network: organises all concepts of the metathesaurus into semantic types and relations (2 semantic types can be linked by several relations):

pharmacologic substance **affects** pathologic function
pharmacologic substance **causes** pathologic function
pharmacologic substance **prevents** pathologic function
...

Semantic Network: example

- CYC: contains common sense knowledge:
  trees are outdoors
  people who died stop buying things …

#$mother :
(#$mother ANIM FEM)
isa: #$FamilyRelationSlot #$BinaryPredicate

See: ontoweb-lt.dfki.de

So, what's an ontology?

- Ontologies are defined as a formal specification of a shared conceptualization
  Borst, 97
- An ontology is a formal theory that constrains the possible conceptualizations of the world
  Guarino, 98

What an ontology is (maybe)

- Community agreement
- Relations between terms
- Pragmatic information
- Common sense knowledge
- Meaning of concepts vs. words: explore language more deeply

Why ontologies?

- ➢ Information retrieval
- ➢ Word Sense Disambiguation
- ➢ Automatic Translation
- ➢ Topic detection
- ➢ Text summarization
- ➢ Indexing
- ➢ Question answering
- ➢ Query improvement
- ➢ Enhance Text Mining

Problem: building an ontology

- Efficiency of the engineering
  - Time
  - Difficulty of the task: ambiguity, completeness
- Agreement of the community

What can be used?

- Texts
- Existing ontologies or core ontologies
- Dictionaries, encyclopediae
- Experts
- Machine Learning and Natural Language Processing tools

What kind of ontology?

- More or less domain specific
- Supervised/unsupervised
- Informal/formal
- For what purpose?
  ⇨determines the granularity, the material, the resources…

Supervised/unsupervised

- One extreme: from scratch
- Other extreme: manual building
- Using a core ontology, structured data…

- Different strategies
- Different tools
- Advantages and inconveniences

Operations on ontologies

- Extraction: building of an ontology
- Pruning: removing what is out of focus; danger: keep the coherence
- Refinement: fine tuning the target (e.g. considering user requirements)
- Merging: mixing of 2 or more similar or overlapping source ontologies
- Alignment: establishing links between 2 source ontologies to allow them to share information
- Evaluation: task-based, necessity of a benchmark!
- …

Components

- Classes of words and concepts
- Relations between concepts
- Axioms defining different kind of constraints
- Instances that can represent specific elements

Relations

- Taxonomic
  hypernym (is a)
    car → vehicle
  hyponym
    fruit → lemon
  events to superordinate
    fly → travel
  events to subtypes
    walk → stroll

Relations

- Meronymic
  From group to members
    team → goalkeeper
    copilot → crew
  From parts to wholes
    book → cover
    wheels → car
  From events to subevents
    snore → sleep

Relations

- Thematic roles

  agent: causer of an event
    "the burglar" broke the window
  experiencer (of an event)
    "the woman" suffers injuries from the car accident
  force: non voluntary causer of an event
    "the earthquake" destroyed several buildings
  theme: participant most directly affected by an event
    the burglar broke "the door"

Relations

- Thematic roles

  instrument (used in an event)
    I've eventually forced the lock "with a screwdriver"
  source: origin of an object of a transfer event
    he's coming "from Norway"
  beneficiary (of an event)
    she's knitting socks "for her grandchildren"

Relations

- Thematic roles can be augmented by the notion of semantic restrictions
- Selectional restrictions: semantic constraint imposed by a lexeme on the concepts that can fill the various arguments roles associated with it
  - "I wanna eat some place that's close to the cinema." "I wanna eat some spicy food."
  - "Which airlines serve Denver?" "Which airlines serve vegetarian meals?"

Part II Text Mining and Natural Language Processing for ontology extraction from text

TM and NLP for ontology extraction from text

- lexical information extraction
- syntactic analysis
- semantic information extraction

Lexical acquisition

collocations

n-grams

Collocations

- A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things
- Technique: count occurrences, rely on frequencies (pb with sparse data)

Mutual information

$$\mathbf{I(x,y) = log[f(x,y)/(f(x)*f(y)]}$$

- extract multiwords units
- group similar collocates or words to identify different meanings of a word
  - bank river
  - bank investment

High similarity?

- Strong ≅ powerful
- I(strong, tea) >> I(powerful, tea)
- I(strong, car) << I(powerful, car)

So…

- Mutual information shows some dissimilarity between "strong" and "powerful", but how can we measure that dissimilarity?
  strong tea vs.* powerful tea

➔ **T-test**

T-test

- Measure of dissimilarity
- Used to differentiate close words (x and y)
- For a set of words, the t-test compares for each word w from this set the probability of having x followed by w to the probability of having y followed by w

Mutual information

I(x,y)	fxy	fx	fy	x	y
10,47	7	7809	28	strong	northerly
9,76	23	7809	151	strong	showings
9,30	7	7809	63	strong	believer
9,04	10	7809	108	strong	currents
8,66	7	1984	388	powerful	legacy
8,58	7	1984	410	powerful	tool
8,35	8	1984	548	powerful	storms
8,32	31	1984	2169	powerful	minority

I(x,y) = log[f(x,y)/(f(x)*f(y)]

T-test

I(strong,w)	t	strong	powerful	w
10,47	1,73	7	0	northerly
9,76	3,12	23	1	showings
9,30	1,73	7	0	believer
9,04	1,22	10	0	currents
I(powerful,w)	t	strong	powerful	w
8,66	-2,53	1	7	legacy
8,58	-2,67	0	7	tool
8,35	-2,33	4	8	storms
8,32	-5,37	3	31	minority

Statistical inference: n-grams

- Consists of taking some data and making some inferences about their distribution: counting words in corpora
- Example: the n-grams model
- The assumption that the probability of a word depends only on the previous word is a Markov assumption.
- Markov models are the class of probabilistic models that assume that we can predict the probability of some future unit without looking too far into the past
  - **A bigram is a first-order Markov model**
  - **A trigram is a second-order Markov model**
  - **…**

Problems

- Wordform / lemma
- Capitalized tokens
- Sparse data
- Deal with huge collections of texts

April 28, 05 37

Example

- “eat” is followed by: on, some, lunch, dinner, at, Indian, today, Thai, breakfast, in, Chinese, Mexican, tomorrow, dessert, British
- “restaurant” is preceded by: Chinese, Mexican, French, Thai, Indian, open, the, a
- Intersection: Chinese, Mexican,Thai, Indian

April 28, 05 38

TM and NLP for ontology extraction from text

- lexical information
- syntactic analysis
- semantic information extraction

April 28, 05 39

Technique: parsing

- Part Of Speech tagging
- Chunking
- Specific relations
- Unsupervised?
- Shallow?
- Efficiency? (resources, processing time)

April 28, 05 40

Example: Shallow Parser

- Tokenizer output
  The patients followed a ‘ healthy ‘ diet and 20% took a high level of physical exercise.
- Tagger output
  The/DT patients/NNS followed/VBD a/DT ‘/” healthy/JJ ‘/” diet/NN and/CC 20/CD %/NN took/VBD a/DT high/JJ level/NN of/IN physical/JJ exercise/NN . /.

April 28, 05 41

Chunker output

[NP The/DT patients/NNS **NP]**
[VP followed/VBD **VP]**
[NP a/DT ‘/” healthy/JJ ‘/” diet/NN **NP]**
and/CC **[NP** 20/CD %/NN **NP]**
[VP took/VBD **VP]**
[NP a/DT high/JJ level/NN **NP]**
{PNP [Prep of/IN Prep] **[NP** physical/JJ exercise/NN **NP] PNP}** . /.

April 28, 05 42

TM and NLP for ontology extraction from text

- lexical information
- syntactic analysis
- semantic information extraction

Techniques

- Selectional restrictions
- Semantic similarity
- Clustering
- Pattern matching

Selectional preferences or restrictions

- The syntactic structure of an expression provides relevant information about the semantic content of that expression
- Most verbs prefer arguments of a particular type

  disease prevented by immunization
  infection prevented by vaccination
  hypothermia prevented by warm clothes

Semantic similarity

- Automatically acquiring a relative measure of how similar a new word is to known words (or how dissimilar) is much easier than determining its meaning.
- Vector space measures: vector similarity
- Add probabilistic measures: refinement

Statistical measures

- Frequency measure:
  F(c,v) = f(c,v) / f(c)+f(v)
- Standard Probability measure:
  P(c|v) = f(c,v) / f(v)
- Hindle Mutual Information measure:
  H(c,v) = log{P(c,v) / [P(v)*P(c)]}
  ► focus on the verb-object cooccurrence

More statistical measures

- Resnik: $R(c,v) = P(c|v) * S_R(v)$
  with $S_R(v) = \sum \{P(c|v) * \log[P(c|v)/ P(c)]\}$
  selectional preference strength
  ► focus on the verb
- Jaccard: J(c,v) = log2 P(c|v) * log2 f(c)/ # c ctx
  with # c ctx = number of contexts of appearance for the compound c
  ► focus on the nominal string

Semantic dissimilarity: Contrastive corpus

- Used to discard
  - general terms
  - unfocused domain terms
- Wall Street Journal vs. Medical corpus

Clustering

- Unsupervised method that consists of partitioning a set of objects into groups or clusters, depending on the similarity between those objects
- Clustering is a way of learning by generalizing.

Clustering

- Generalizing: assumption that an environment that is correct for one member of the cluster is also correct for the other members of the cluster
- Example: preposition to use with “Friday” ?
  1.Existence of a cluster “ Monday, Sunday, Friday”
  2. Presence of the expression “on Monday”
  3. Choice of the preposition “on” for “Friday”

Types of clustering

- Hierarchical: each node stands for a subclass of its mother’s node; the leaves of the tree are the single objects of the clustered sets
- Non hierarchical or flat: relations between clusters are often undetermined
- Hard assignment: each object is assigned to one and only one cluster
- Soft assignment allows degrees of membership and membership in multiple clusters (uncertainty)
- Disjunctive clustering: “true” multiple assignment

Hierarchical

- Bottom-up (agglomerative): starting with each objet as a cluster and grouping the most similar ones
- Top-down (divisive clustering): all objects are put in one cluster and the cluster is divided into smaller clusters (use of dissimilarity measures)

Example bottom-up

- Three of the 10000 clusters found by Brown et al, (1992), using a bigram model and a clustering algorithm that decreases perplexity:
  - plan, letter, request, memo, case, question, charge, statement, draft
  - day, year, week, month, quarter, half
  - evaluation, assessment, analysis, understanding, opinion, conversation, discussion

Non hierarchical

- Often starts with a partition based on randomly selected seeds (one seed per cluster) and then refine this initial partition
- Several passes are often necessary. When to stop? You need to have a measure of goodness and you go on as long as this measure is increasing enough

Examples

- AutoClass (Minimum Description Length): the measure of goodness captures both how well the objects fit into the clusters and how many clusters there are. A high number of clusters is penalized.
- EM alorithm
- K-means
- …

Pattern matching / Association rules

Pattern matching consists of finding patterns in texts that induce a relation between words, and generalizing these patterns to build relations between concepts

Srikant and Agrawal algorithm

This algorithm computes association rules Xk ⇒ Yk, such that measures for support and confidence exceed user-defined thresholds.
Support of a rule Xk ⇒ Yk is the percentage of transactions that contain Xk U Yk as a subset
Confidence is defined as the percentage of transactions that Yk is seen when Xk appears in a transaction.

Example

- Finding associations that occur between items, e.g. supermarket products, in a set of transactions, e.g. customers' purchases.
- Generalization:
  "snacks are purchased with drinks" is a generalization of
  "chips are purchased with bier" or
  "peanuts are purchased with soda"

References

- Manning and Schutze, "Foundations of Statistical natural Language Processing"
- Mitchell, "Machine Learning"
- Jurafsky and Martin, "Speech and Language Processing"
- Church et al., "Using Statistics in Lexical Analysis". In Lexical Acquisition (ed. Uri Zernik)

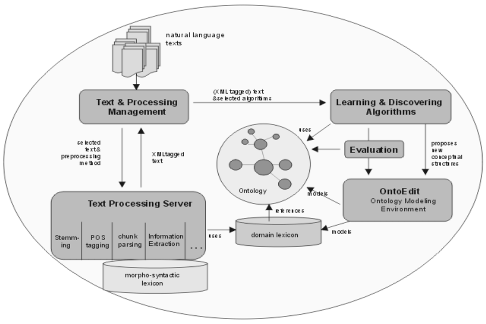
Part III: Ontology Building Systems

1. TextToOnto (AIFB, Karlsruhe)
2. CORPORUM-OntoBuilder (Ontoknowledge project)
3. OntoLearn
4. Mumis (European project)
5. OntoBasis (CNTS)

1. Text To Onto

This system supports semi-automatic creation of ontologies by applying text mining algorithms.

The Text-To-Onto system



Semi-automatic ontology engineering

- Generic core ontology used as a top level structure
- Domain specific concepts acquired and classified from a dictionary
- Shallow text processing
- Term frequencies retrieved from texts
- Pattern matching
- Help from an expert to remove concepts unspecific to the domain

Learning and discovering algorithms

- The term extraction algorithm extracts from texts a set of terms that can potentially be included in the ontology as concepts.
- The rules extraction algorithm extracts potential taxonomic and non-taxonomic relationships between existing ontology concepts. Two distinct algorithms:
  the regular expression-based pattern matching algorithm mines a concept taxonomy from a dictionary
  the learning algorithm for discovering generalized association rules analyses the text for non-taxonomic relations
- The ontology pruning algorithm extracts from a set of texts the set of concepts that may potentially be removed from the ontology.

Learning algorithm

- Text corpus for tourist information (in German), that describes locations, accomodations, administrative information…
- Example: Alle Zimmer sind mit TV, Telefon, Modem und Minibar ausgestattet. (All rooms have TV, telephone, modem and minibar.)
- Dependency relations output for that sentence: Zimmer – TV (room – television)

Example

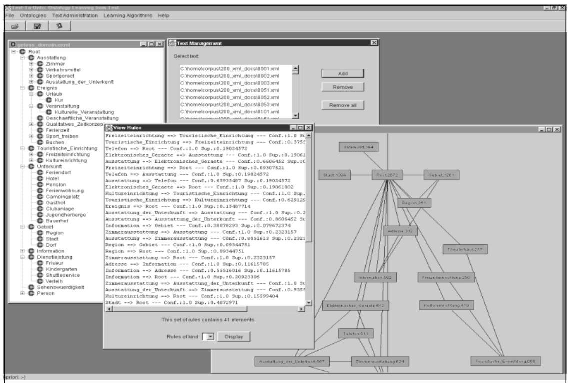
- Tourist information text corpus
- Concepts pairs derived from the text:
  area – hotel
  hairdresser – hotel
  balcony – access
  room – television

- Domain taxonomy

Root
furnishing
accomodation area
hotel region city

- Discovered relations

	Support	Confidence
(area, accomodation)	0.38	0.04
~~(area, hotel)~~	0.1	0.03
(room, furnishing)	0.39	0.03
~~(room, television)~~	0.29	0.02
(accomodation, address)	0.34	0.05
(restaurant, accomodation)	0.33	0.02



Ontology: example

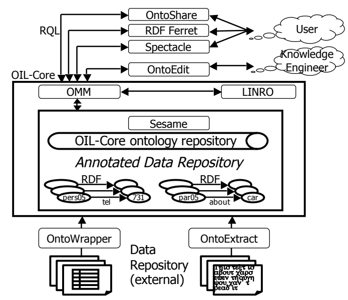
```
- <rdfs:Class rdf:about="test:cat">
  <rdfs:subClassOf rdf:resource="test:animal" />
  </rdfs:Class>
- <rdfs:Class rdf:about="test:persian_cat">
  <rdfs:subClassOf rdf:resource="test:cat" />
  </rdfs:Class>
<!-- properties of cars and cats -->
- <rdf:Property rdf:about="test:color">
  <rdfs:domain rdf:resource="test:car" />
  <rdfs:domain rdf:resource="test:cat" />
  </rdf:Property>
<!-- properties between cars and cats -->
- <rdf:Property rdf:about="test:runs_over">
  <rdfs:domain rdf:resource="test:car" />
  <rdfs:range rdf:resource="test:cat" />
  </rdf:Property>
```

http://kaon.semanticweb.org/frontpage

2. Ontoknowledge

Content-driven Knowledge-Management through Evolving Ontologies

The overall architecture and language



OntoBuilder

- Ontowrapper: structured documents (names, telephone numbers…)
- OntoExtract: unstructured documents
  - provide initial ontologies through semantic analysis of the content of web pages
  - refine existing ontologies (key words, clustering…)

OntoWrapper

• Deals with data in “regular” pages

• Uses personal “extraction rules”

• Outputs instantiated schemata

April 28, 05 73

OntoExtract

Taking a single text or document as input, *OntoExtract* retrieves a document specific light-weight ontology from it.

Ontologies extracted by *OntoExtract* are basically taxonomies that represent *classes, subclasses* and *instances*.

April 28, 05 74

OntoExtract: Why?

• concept extraction
• relations extraction
• semantic discourse representation
• ontology generation
• part of document annotations
• document retrieval
• document summarising
• ...

April 28, 05 75

OntoExtract: How?

Extraction Technology based on
– *tokeniser*
– *morphologic analysis*
– *lexical analysis*
– *syntactic/semantic analysis*
– *concept generation*
– *relationships*

April 28, 05 76

OntoExtract

• ***learning initial ontologies***
-> propose networked structure

• ***refining ontologies***
-> add concepts to existing onto’s
-> add relations “across” boundaries

April 28, 05 77

OntoExtract

- *Classes*, described in the text which is analysed.
- *Subclasses*, classes can also be defined as subclass of other classes if evidence is found that a class is indeed a subclass of another class.
- *Facts/instances:* Class definitions do not contain properties. As properties of classes are found, they will be defined as properties of an instance of that particular class.

The representation is based on relations between classes based on semantic information extracted.

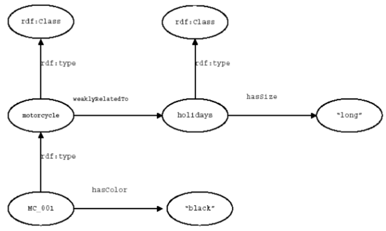
April 28, 05 78

Example

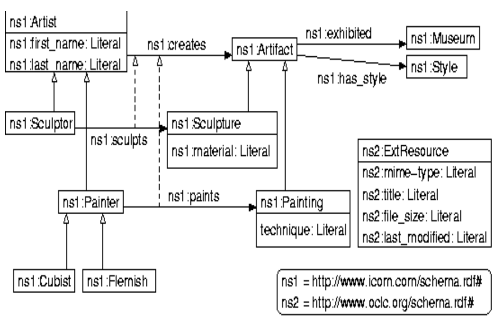
<rdfs:Class rdf:ID="news_service">
<rdfs:subClassOf rdf:resource="#service"/>
</rdfs:Class>
<news_service rdf:ID="news_service_001">

<hasSomeProperty>financial</hasSomeProperty>
</news_service>

Ontology: example



Museum repository



Query example

http://sesame.aidministrator.nl/publications/rql-tutorial.html#N366
http://sesame.aidministrator.nl/sesame/actionFrameset.jsp?repository=museum

select X, $X, Y from {X : $X} cult:paints {Y} using namespace cult = http://www.icom.com/schema.rdf#

select X, Z, Y from {X} rdf:type {Z}, {X} cult:paints {Y} using namespace rdf = http://www.w3.org/1999/02/22-rdf-syntax-ns# , cult = http://www.icom.com/schema.rdf#

select X, Y from {X : cult:Cubist } cult:paints {Y} using namespace cult = http://www.icom.com/schema.rdf#

select X, $X, Y from {X : $X} cult:last_name {Y} where ($X <= cult:Painter and Y like "P*") or ($X <= cult:Sculptor and not Y like "B*") using namespace cult = http://www.icom.com/schema.rdf#

select PAINTER, PAINTING, TECH from {PAINTER} cult:paints {PAINTING}. cult:technique {TECH} using namespace cult = http://www.icom.com/schema.rdf#

Query example

select PAINTER, PAINTING, TECH from {PAINTER} cult:paints {PAINTING}. cult:technique {TECH} using namespace cult = http://www.icom.com/schema.rdf#

Query results: PAINTER PAINTING TECH

http://www.european-history.com/picasso.html http://www.european-history.com/jpg/guernica03.jpg "oil on canvas"@en

http://www.european-history.com/picasso.html http://www.museum.es/woman.qti "oil on canvas"@en

http://www.european-history.com/rembrandt.html http://www.artchive.com/rembrandt/artist_at_his_easel.jpg "oil on canvas"@en

http://www.european- history.com/rembrandt.html http://www.artchive.com/rembrandt/abraham.jpg "oil on canvas"@en

http://www.european-history.com/goya.html http://192.41.13.240/artchive/graphics/saturn_zoom1.jpg "wall painting (oil)"@en

5 results found in 323 ms.

http://www.ontoknowledge.org

OntoLearn

An infrastructure for automated ontology learning from domain text.

Semantic interpretation

- Identifying the right senses (concepts) for complex domain term components and the semantic relations between them.
- use of WordNet and SemCor
- creation of Semantic Nets
- use of Machine Learned Rule Base
- Domain concept forest

Ontology Integration

- from a core domain ontology or from WordNet
- Applied to multiword term translation

http://www.ontolearn.de

4. MUMIS

Goal: to develop basic technology for automatic indexing of multimedia programme material

MUMIS

- Use data from different media sources (documents, radio and television programmes) to build a specialised set of lexica and an ontology for the selected domain (soccer).
- Access to textual and especially acoustic material in the three languages English, Dutch, and German

MUMIS

- Domain: soccer
- Developement of an ontology and a multi-language lexica for this domain
- Query: "give me all goals Uwe Seeler shot by head during the last 5 minutes of a game" (formal query interface)
- Answer: a selection of events represented by keyframes

Information Extraction

- Natural Language Processing (Information Extraction)
  - Analyse all available textual documents (newspapers, speech transcripts, tickers, formal texts ...), identify and extract interesting entities, relations and events
- The relevant information is typically represented in form of predefined “templates”, which are filled by means of Natural Language analysis
- IE combines here pattern matching, shallow NLP and domain knowledge
- Cross-document co-reference resolution

IE DATA

Ticker
24 Scholes beats Jens Jeremies wonderfully, dragging the ball around and past the Bayern Munich man. He then finds Michael Owen on the right wing, but Owen's cross is poor.

TV report
Scholes
Past Jeremies
Owen

Newspaper
Owen header pushed onto the post
Deisler brought the German supporters to their feet with a buccaneering run down the right. Moments later Dietmar Hamann managed the first shot on target but it was straight at David Seaman. Mehmet Scholl should have done better after getting goalside of Phil Neville inside the area from Jens Jeremies' astute pass but he scuffed his shot.

Formal text

Schoten op doel	4 4
Schoten naast doel	6 7
Overtredingen	23 15
Gele kaarten	1 1
Rode kaarten	0 1
Hoekschoppen	3 5
Buitenspel	4 1

IE Techniques & resources

- Tokenisation
- Lemmatisation
- POS + morphology
- Named Entities
- Shallow parsing
- Co-reference resolution
- Template filling

24 Scholes beats Jens Jeremies wonderfully, dragging the ball around and past the Bayern Munich man. He then finds Michael Owen on the right wing, but Owen's cross is poor.

24	time
Scholes	player
beat	
Jens Jeremies	player
wonderfull	
,	
…	

IE subtasks

- 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.
- Template Element task (TE): Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text.

Terms as descriptors and terms for NE task

Team: *Titelverteidiger* Brasilien, den respektlosen *Außenseiter* Schottland

Trainer: Schottlands *Trainer* Brown, *Kapitän* Hendry seinen *Keeper* Leighton

Time: *in der 73. Minute*, *nach gerade einmal 3:50 Minuten*, von Roberto Carlos *(16.)*, *nach einer knappen halben Stunde*,

IE subtasks

- Template Relation task (TR): Extract relational information on employee_of, manufacture_of, location_of relations etc. (TR expresses domain-independent relationships).

Opponents: Brasilien *besiegt* Schottland, *feierte* der Top-Favorit

Trainer_of: Schottland*s* Trainer Brown

IE subtasks

- Scenario Template task (ST): Extract pre-specified event information and relate the event information to particular organization, person, or artifact entities (ST identifies domain and task specific entities and relations).

Foul: als er den durchlaufenden Gallacher im Strafraum allzu energisch am Trikot *zog*

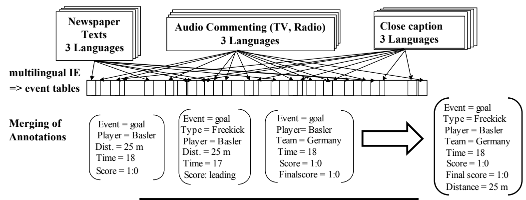
Substitution: und mußte in der 59. Minute für Crespo *Platz machen…*

IE subtasks

- Co-reference task (CO): Capture information on co-referring expressions, i.e. all mentions of a given entity, including those marked in NE and TE.

April 28, 05 97

Off-line Task



Events indexed in video recording

•Freekick	•Goal	•Pass	•Defense
•17 min	•18 min	•24 min	•28min
	•1:0		
•Foul	•Freekick		•Dribbling
•Neville	•Basler	•Matthäus	•Campbell
•Basler			•Scholl
•25 m	•25 m	•60 m	

April 28, 05 98

On-line task

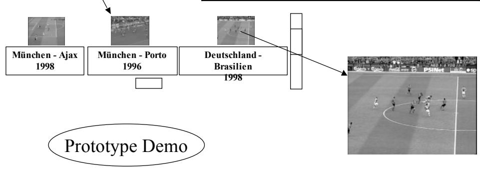
- Searching and Displaying
- Search for interesting events with formal queries
  Give me all goals from Overmars shot with his head in 1. Half.
  Event=Goal; Player=Overmars; Time<=45; Previous-Event=Headball
- Indicate hits by thumbnails & let user select scene
- Play scene via the Internet & allow scrolling etc
- User Guidance (Lexica and Ontology)

April 28, 05 99

On-line task

Knowledge Guided User Interface & Search Engine

•Freekick	•Goal	•Pass	•Defense
•17 min	•18 min	•24 min	•28min
	•1:0		
•Foul	•Freekick		•Dribbling
•Neville	•Basler	•Matthäus	•Campbell
•Basler			•Scholl
•25 m	•25 m	•60 m	



Prototype Demo

April 28, 05 100

5. OntoBasis

Elaboration and adaptation of semantic knowledge extraction tools for the building of specific domain ontology

April 28, 05 101

Unsupervised learning

[NP1Subject The/DT Sarsen/NNS Circle/NNP **NP1Subject] [VP1** is/VBZ **VP1]…**

mutation in gene

catalytic_subunit of DNA_polymerase

raw text
↓ shallow parser
parsed text
↓ pattern matching
relations
↓ statistics
relevant relations
↓ evaluation
initiation of an ontology

April 28, 05 102

Material

- Stonehenge corpus, 4K words, rewritten
- Extraction of semantic relations using pattern matching and statistical measures
- Focus on "part of" and spatial relations, dimensions, positions…

Stonehenge corpus

- Description of the megalithic ruin
  The trilithons are ten upright stones
  The Sarsen heel stone is 16 feet high.
  The bluestones are arranged into a horseshoe shape inside the trilithon horseshoe.

Syntactic analysis

The Sarsen Circle is about 108 feet in diameter .

The/DT Sarsen/NNS Circle/NNP is/VBZ about/IN
108/DT feet/NNS in/IN diameter/NN ./.

[NP The/DT Sarsen/NNS Circle/NNP **NP]**
[VP is/VBZ**VP]**
[NP about/IN 108/DT feet/NNS **NP]**
[PP in/IN **PP]** **[NP** diameter/NN **NP]** ./.

[NP1Subject The/DT Sarsen/NNS Circle/NNP **NP1Subject]**
[VP1 is/VBZ **VP1]**
[NP about/IN 108/DT feet/NNS **NP]**
{PNP [PP in/IN PP] [NP diameter/NN NP] **PNP}** ./.

Pattern matching

- Selection of the syntactic structures
  Nominal String – Preposition – Nominal String
  Ns-Prep-Ns
  [a Ns is a string of adjectives and nouns, ending up with the head noun of the noun phrase]

  Edman_degradation of intact_protein
  beta-oxidation of fatty_acid
  56_Aubrey_hole inside circle

Selection

- Nominal Strings filtering using a statistical measure: the measure is high when the prepositional structure is coherent
- We select the N most relevant structures

$$\frac{\frac{\#NS1\text{-}P\text{-}NS2}{Min(\#NS1,\#NS2)}}{\frac{\#NS1\text{-}P}{\#NS1} + \frac{\#P\text{-}NS2}{\#NS2}}$$

Pattern matching

- Syntactic structures Subject-Verb-Direct Object or "lexons"

amino_acid_sequence show Bacillus_subtilis
nucleotide_sequencing reveal heterozygosity
Aubrey_Holes are inside bank

Combination

- We consider the N prepositional structures with the highest rate selected previously
- We elect the structures Sub-Vb-Obj where the Subject and the Object both appear among those N structures

Examples

- "part of" basic relations
  bottom of stone
  shape of stone
  block of sandstone
- spatial relations
  ring of bluestones
  center of circle
  sandstone on Marlborough Downs
  Preseli Mountain in Pembrokeshire
- disposition of the stones
  Bluestone circle outside Trilithon horseshoe
  Bluestone circle inside Sarsen Circle
  Bluestone circle is added outside Trilithon horseshoe
  Slaughter Stone is made of sarsen
  100 foot diameter circle of 30 sarsen stone

Wrong relations

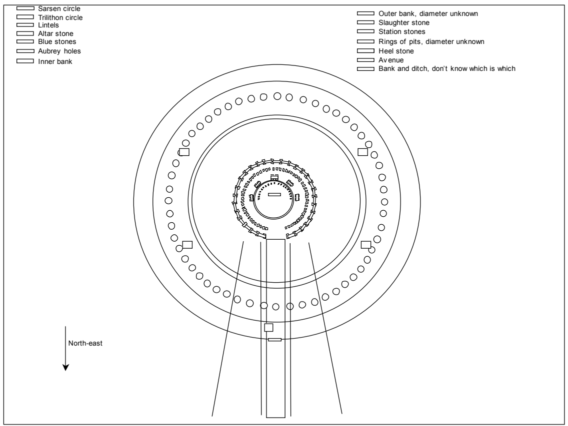
Altar Stone is in front
Heel stone leans of vertical
Sarsen block are 1.4 metre
Stonehenge is of 35 foot
heel stone is from ring
120 foot from ring
Two of Station Stone
central part of monument
rectangle to midsummer sunrise line of monument
...

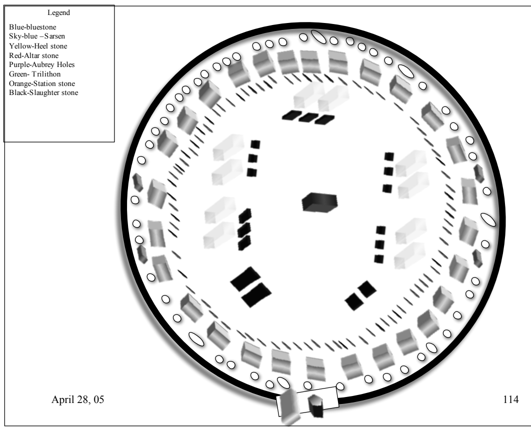
- Incomplete
- Uninformative
- Irrelevant

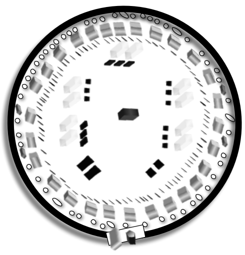
Correct relations we didn't use

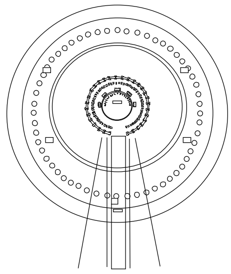
Aubrey Holes vary from 2 to 4 foot in depth
8-ton Heel Stone is on main axis at focus
Sarsen stone are from Marlborough Down
Stonehenge stands on open downland of Salisbury Plain
bluestone came from Preselus Mountain in southwestern Wale
monument comprises of several concentric stone arrangement
Heel Stone is surrounded by circular ditch
third trilithon stone bears of distinguished human head
carving on twelve stone
trilithon linteled of large sarsen stone
Three Trilithon are now complete with lintel
…

- Provenance - locations
- Sizes - weight
- Details (carvings)









April 28, 05 115

Results

- What we get:
  positions amounts
  sizes weights
  composition (shape)
- Double checking of some information possible due to different descriptions and/or different patterns relevant on the same phrase
- World knowledge lacking
- Information uncomplete

April 28, 05 116

WebSites

- http://kaon.semanticweb.org/frontpage
- http://www.ontoknowledge.org
- http://www.ontolearn.de

- http://wise.vub.ac.be/ontobasis
- http://www.cnts.ua.ac.be/cgi-bin/ontobasis

April 28, 05 117