Knowledge Representation and Extraction for Business Intelligence

Thierry Declerck (DFKI), Horacio Saggion (University of Sheffield), Marcus Spies (STI University of Innsbruck)
Notes

- Contributors
  - Christian Leibold
  - Hans-Ulrich Krieger
  - Bernd Kiefer

- Slides and updates at:
  - http://www.gate.ac.uk/conferences/iswc08-tutorial
Main Objectives of the MUSING Project

- Creation of the next generation of industrial analysis: the semantic-based Business Intelligence;
  - Development and validation of BI solutions with emphasis on Credit Risk Management (Basel II and beyond);
  - Development and validation of semantic-based internationalisation platforms;
  - Development and validation of semantic-driven knowledge systems for IT-OpR measurement and mitigation tools, with particular reference to operational risks/business continuity issues faced by IT-intensive organisations;

- Validation of the research and technological development results in those domains with high societal impact. Exploitation of the multi-industry potential.
Main Research and Development objectives

- Knowledge management & reasoning
- Natural language processing & semantic web
- Representation of temporal information
- European Internationalisation policies
- (Bayesian) integration of qualitative and quantitative knowledge elements
- Integration of the various scientific communities involved in MUSING
- Contributions to standards
General overview of semantic technologies in MUSING
MUSING Ontologies

General Ontologies
- Company
- Industry Sector
- BACH
- XBRL
- Risk

Adapted Proton Ontologies
- System
- Top
- Upper
- Extension

Temporal Ontologies
- 4D
- Time

Financial Services
- Qualitative Analysis

Internationalisation
- Indicator
- Region
- Employment
- Real Estate

Operational Risk Management
- Operational Risk
- Loss Event
Data Sources in MUSING

- Data sources are provided by MUSING partners and include balance sheets, company profiles, press data, web data, etc. (some private data)
  - Il Sole 24 ORE, CreditReform data
  - Companies’ web pages (main, “about us”, “contact us”, etc.)
  - Wikipedia, CIA Fact Book, etc.

- Ontology is manually developed through interaction with domain experts and ontology curators
  - It extends the PROTON ontology and covers the financial, international, and IT operative risk domain
Processing Structured and Unstructured Data

- Ontology-driven analysis of both structured and unstructured textual data
  - Structured Data
    - Profit & Loss tables (which are structured but not normalized: extracting from the tables the data (terms, values, dates, currency, etc.) and map them into a normalized representation in XBRL, the eXtensible Business Reporting Language.
    - Company Profiles and International Reports, which give detailed information about company (name, address, trade register, share holders, management, number of employees etc.)
  - Unstructured Data
    - Annexes to Annual Reports, On-Line financial articles, questionnaire to credit institutions etc.

- The Challenge: Merging data and information extracted from various types of documents (structured and unstructured), using a combination of Ontologies/Knowledge Bases, linguistic analysis and statistical models
Examples of the processing of Structured data sources

- The PDFtoXBRL tools
  - Extract financial tables from PDF documents (Annual reports of companies)
  - Reconstruct a tabellar representation of the information contained in the tables (dates, amount, financial terms etc.) and annotate those with the corresponding semantics
  - Map to a standardized representation (for example GAAP in XBRL).
  - Good quality so far: depending on the quality of the processable input document: 75% up to 95% F-Measure.
Ontology-Based Information Extraction in MUSING

DATA SOURCE PROVIDER -> DOCUMENT COLLECTOR -> DOCUMENT

MUSING DATA REPOSITORY -> DOCUMENT -> DOCUMENT

DOMAIN EXPERT -> ONTOLOGY-BASED DOCUMENT ANNOTATION

ONTOLEGY-POPULATION -> INSTANCES & RELATIONS

MUSING ONTOLOGY CURATOR

MUSING APPLICATION

ECONOMIC INDICATORS
REGION SELECTION MODEL
COMPANY INFORMATION

REGION RANK

USER INPUT

USER

MUSING DATA REPOSITORY

KNOWLEDGE BASE
Ontology Extension/Extraction

- Manual expert-based ontology generation is very time consuming. How to partially automatize this task?
  
  ✓ Extracting from documents possible candidates for ontology classes and relations, using a combination of linguistic analysis, semantic annotation and statistical models. A first "shallow" prototype has been implemented.

  ✓ So for example, in XBRL (2.0) the values for members of boards are of string-type (ordered in a flat list). From textual analysis of Annual reports we could extract a further possible hierarchy within the "members of boards", and suggest a more fine-grained representation of the information associated with the members of boards.
MUSING in action: Financial Risk Management (FRM)

Third party customer (e.g.: SME that wants to perform a Basel II self-assessment)

MUSING services in Financial Risk Management

Through web services

MUSING FRM customer (e.g.: financial institution that wants to acquire prospect lists)

Client’s proprietary info (structured & unstructured)

NOTE: In MUSING FRM is essentially focusing on credit merit assessment, keeping a close eye on all Basel II implications

External Info (web etc...)

MUSING info & ontologies
Expected Impact of MUSING in FRM

- Improving the access to credit for SMEs in Basel II scenario and beyond
  - total cost for Financial Institutions to adopt Basel II-compliant risk mgnt systems in the EU will be between 20bn and 30bn between 2002-2006 (Pricewaterhouse Coopers’ Study)
- Automating banking procedures related to credit issuing workflow
- Improving Business Reporting through Standardisation and Ontologisation of existing taxonomies (for example XBRL)
- Supporting Professionals’ daily work
A scenario in the FRM domain

- Support the new way of working introduced by Basel II, that involves feeding the internal rating systems of financial institutions
- Test the ability of the MUSING solutions to automatically extract information from Balance Sheets (both P&L, A&L and their annexes – e.g. “Nota Integrativa”, for the Italian specific case)
- The scenario:
  - Upload a balance sheet document (in PDF)
  - Transform the content of the tables into XBRL (eXtensible Business Reporting Language)
  - Submit to the operator for checking, and include in her/his workflow
  - Present to the operator direct links to the relevant parts of the NI that are giving more information to the specific XBRL item
  - Integrate the feedback of the operator (corrected XBRL document) into the extraction mechanism
Graphical View of the Scenario
Structured Data in the Scenario

- Profit & Loss tables etc. are structured but not normalized.
  - First processing step consists in automatically extracting from the balance tables the data (terms, values, dates, currency, etc.) and map them into a XBRL representation (the MUSING PDF2XBRL tools)
Unstructured Data in the Scenario

- Annexes to Italian Annual Reports - Example of free text in the unstructured part of the annex
  - “Le immobilizzazioni materiali sono iscritte al costo di acquisto o di produzione al netto dei relativi fondi di ammortamento, inclusi tutti i costi e gli oneri accessori di diretta imputazione, dei costi indiretti inerenti la produzione interna, nonché degli oneri relativi al finanziamento della fabbricazione interna sostenuti nel periodo di fabbricazione e fino al momento nel quale il bene può essere utilizzato. ...”

- Linguistic and semantic analysis of such textual documents results in Semantic metadata that enrich the original document.

- Out of this kind of text, definitions can be automatically extracted but also (semantic) relations, like the one between immobilizzazioni materiali and costo di acquisto o di produzione”, etc.
Automatic Links between XBRL Positions and the Nota Integrativa

- Aligning the normalized quantitative information in the financial tables with the relevant text parts in the annex Nota Integrativa, supporting the work of the operator (also towards a XBRL normalization of the unstructured parts of the Nota Integrativa)
A Proposal for Temporal Representation and Reasoning in the MUSING Project

Hans-Ulrich Krieger, Bernd Kiefer, Thierry Declerck (DFKI GmbH)
Motivation: Example 1

Dieter Zetsche ist der neue Vorstandsvorsitzende von DaimlerChrysler.
<dc,rdf:type,Company>
<dz,rdf:type,Person>
<dc,hasCeo,dz>

**problem:** *synchronic* representation refers to one point in time *(which point?)*
Motivation: Example 2

most relationships are \textit{diachronic}, i.e., they vary with time

\textit{Jürgen Schrempp} gibt bekannt, daß er zum 31. Dezember 2005 als \textit{Vorstandsvorsitzender} von \textit{DaimlerChrysler} \textit{ausscheiden} wird.

\[ t = 2005-12-31: \langle js,\text{resignsFrom},dc\rangle \]
\[ ? \leq t \leq 2005-12-31: \langle js,\text{ceoOf},dc\rangle \]
Example 2, cont.

1995 gab Edzard Reuter den Vorstandsvorsitz der Daimler Benz AG an Schrempp ab.

\[1995 \leq t \leq ?: \langle s, \text{ceoOf,db} \rangle\]

need to identify entities that are referred to by different referential expressions (e.g., Jürgen Schrempp, Schrempp, der Vorstandsvorsitzende von DC, er)
Example 2, cont.

Er ist unter anderem bei der Allianz AG und bei Vodafone Mitglied des Aufsichtsrats.

$t1 \leq t \leq t2$: <e1,memberOfSupBoard,a>
$t3 \leq t \leq t4$: <e1,memberOfSupBoard,v>
<e1,owl:sameAs,js>

heuristics (for present tense): take date of document (= t) into account to have at least a safe time point where the above proposition holds: $t1 = t2 = t3 = t4 = t$
Examples From MUSING: Changing Relationships

most (all?) relations change over time

- name of a company
- CEO of a company
- company address
- win & loss of a company
- number of employees
- members of management board
- ...

Diachronic Identity

need to identify individuals that are different at different times, but refer to the same entity

- **observation 1**: value of a property is only valid within a certain time interval
  (example 2: CEOship)
Diachronic Identity

- **Observation 2**: Property must not hold for each subinterval (aka subinterval inheritance)
  
  - Die Deutsche Bank steigerte ihren Ergebnis vor Steuern in 2005 um 58%. (no constant raise of 58% over whole year)
  
  - Yesterday we drove west. (we mostly drove west)
DI: Endurants vs. Perdurants

- 3D/endurantist view
  - distinction between endurants & occurrants
    - endurants: wholly present
    - occurrants: have temporal parts
    - DI of endurants: essential properties must always hold
DI: Endurants vs. Perdurants

- **4D/perdurantist view**
  - all entities (simple event ... lifetime universe) exist for some period of time
  - *spacetime worms* (Sider 1997) = 4D trajectory

- **MUSING: adopt perdurantist view (time only)**
  - associate entity with all its temporal parts
Technical Approaches To DI

- equip relation with a temporal argument
  - temporal data bases, logic programming
  - $\text{hasCeo}(\text{dc}, \text{js}) \rightarrow \text{hasCeo}(\text{dc}, \text{js}, t)$

- apply meta-logical predicate $\text{hold}$
  - McCarthy&Hayes, Allen, KIF
  - $\text{hold} (\text{hasCeo} (\text{dc}, \text{js}), t)$
Approaches To DL, cont.

- **reification**
  - ✔ RDF
  - ✔ wrap original arguments in a new object
  - ✔ introduce new class, say CEO, for companies & persons: 
    \[ \text{hasCeo}(dc, js) \rightarrow \text{hasCeo}(dc, js, t) \]
  - ✔ \( \text{type}(cp, CEO) \land \text{hasTemporalExtension}(cp, t) \land \text{company}(cp, dc) \land \text{person}(cp, js) \)
Reification/Wrapping & OWL

- need to introduce a new class & accessor for each property that changes over time
- some forms of built-in OWL reasoning no longer possible (Welty et al. 2005)

- reasoning/querying more complex
  ✓ example: return all CEOs of DC
  ✓ (S) SELECT ?comp WHERE {dc hasCeo ?comp}
  ✓ (D) SELECT ?comp WHERE {?ceo rdf:type CEO.
  ?ceo company dc. ?ceo person...}
DL/OWL and DI

- DL/OWL supports
  - binary (and unary) relations only
    - `hasCeo(dc,js,t)` – does **not** work!
  - no complex relation arguments
    - `hold(hasCeo(dc,js,t))` – does **not** work!
DL/OWL and DI (cont.)

- so, use reification – NO!
  - at least not on the original arguments
  - distinguished first argument of a relation: domain
  - associate individual in 1st place with all its temporal facts/parts
    - introduce a time slice (remember spacetime worms)
    - TS = co-occurring information holds for same time period
    - perdurant (a spacetime worm) = container of time slices
Ontology – Structure

**Perdurant**: hasTimeSlice ` timeSliceOf`, plus temporary-constant properties

**TimeSlice**: timeSliceOf, hasTemporalEntity, plus domain-dependent properties

**TemporalEntity**: qualifier (absolute, every, ...)

- **Instant**
  - NegativeInfinity
  - PositiveInfinity
  - ProperInstantYear: $\leq 1$ year
  - ProperInstantMonth: plus $\leq 1$ month
  - ProperInstantDay: plus $\leq 1$ day

- **Interval**: $\leq 1$ begins, $\leq 1$ ends
  - Forever
  - UndefinedInterval:
    - OpenLeftInterval: $= 1$ ends
    - ClosedInterval
    - OpenRightInterval: $= 1$ begins
    - ClosedInterval
Ontology – Structure, cont.

ClosedInterval, OpenLeftInterval, and OpenRightInterval

\(\forall\text{begins. ProperInstantYear} \cup \forall\text{ends. ProperInstantYear}\)

Day, \(\forall\text{begins. ProperInstantDay} \cup \forall\text{ends. ProperInstantDay}\)

Monday, Thursday, ...

SpecialDay

Christmas, ...

NewYearsEve, \(\forall\text{begins. (9month.}{12}\cup 9\text{day.}{31})\)

Month

January, February28, February29, ...

Quarter

FirstQuarter, SecondQuarter, ...

Season

Spring, Summer, ...
Ontology – Remarks

- Intervals must **not** be convex (might contain holes)
  - Example: Yesterday, we drove west
  - Car might have even stopped (& mostly drove west)

- **No** distinction between open & closed intervals
  - I.e., $<s,t>$ always meets $<t,u>$ ($< \in \{(, [], > \in \{(), \}])$
  - More subtle distinction probably not needed in MUSING
Ontology – Remarks

- time slice of a perdurant either refers to interval or instant
  - On January 1, 2002 (00:00:0), the Euro was officially introduced.

- granularity of an instant can be arbitrarily detailed
  - properties on ProperInstantXXX: year, month, day, hour, ...
  - determines whether instant/interval is partially/fully specified
  - alternative to subtyping: cardinality constraints
Consequences of Using OWL

- binary OWL properties can NOT be extended by further time arguments
  
  [ should we move to a different language, e.g., F-logic ]

- wrap property value plus temporal information in a time slice object

- what had originally been an entity (e.g., person, company) now becomes a time slice

- access to time slices of a perdurant via hasTimeSlice property
Wrong Representation

- person $p$ was CEO for two companies $c_1, c_2$
  - $[s_1, s_2] : \text{ceoOf}(p, c_1)$
  - $[t_1, t_2] : \text{ceoOf}(p, c_2)$
- wrong associations, e.g., $[s_1, s_2] : \text{ceoOf}(p, c_2)$
Right Representation

person p1, p2 & company c1, c2 become time slices; introduce new perdurant P
From Entities to Time Slices

- what was an entity now becomes a time slice
  - do not reduplicate PROTON's psys:Entity class hierarchy on the perdurant side
  - example: ptop:Person represents a time slice of a perdurant that acts as a person
  - move time-varying information into a perdurant's TS
From Entities to Time Slices (cont.)

- move temporal-constant information to the perdurant
- a perdurant might have TSs of different types
- approach makes it easy to accommodate 3D space
Grounding in OWL-Time & PROTON

- TemporalEntity, Instant & Interval and begins & ends do exist in OWL-Time
- delete subclass ptop:TimeInterval of class ptop:Happening
- remove ptop:startTime and ptop:endTime from ptop:Happening
- delete subclass pup:TemporalAbstraction of class ptop:Abstract
- psys:Entity ≡ time:TimeSlice
  ✓ subclasses: Abstract, Happening, Object
Temporal Abstractions, e.g., pupp:CalendarMonth, are viewed as temporal abstractions

- not equipped with properties that deal with temporal extension, such as startTime, endTime
- we view them as potentially underspecified periods of time
- CalendarMonth "inherits" properties from superclass ptop:Entity, such as ptop:partOf or ptop:locatedIn
- temporal abstraction hierarchy somewhat arbitrary
  - day of month is a temporal abstraction
  - a river as such is NOT a locative abstraction (there is no such class), but instead a subclass of ptop:Object (very concrete)
Removing Time from PROTON, cont.

- `ptop:startTime` and `ptop:endTime` are defined on `ptop:Happening` (not on `ptop:TimeInterval`)
- effect: instances from `ptop:Object`, e.g., from classes Company or Person, can not be given a temporal extend
- no distinction between instant and interval in PROTON (Instant not expressible as a subclass of `TimeInterval` in TBOX: would require role-value map)
- nearly every property defined on `psys:Entity` might change over time, thus `Entity ≡ TimeSlice`

p1 and p2: time slices of perdurant js (entity Jürgen Schrempp)
c1 and c2: time slices of perdurant dc (entity DaimlerChrysler)

---

OWLIM rule to "close" intervals

BUT: begins & ends are functional props

SELECT min(?begins) max(?ends)
WHERE { mus:js time:hasTimeSlice ?ts.
  ?ts time:hasTemporalEntity ?int.
  ?int time:ends ?ends.}

**effect:** min/max treatment can handle different time slices of same person for ceoOf relation, assuming (heuristics) that ceoOf lasts between min and max

**problem:** SPARQL does not come up with min/max (but SQL)

**general rule:** abstract from a specific person and a specific relation; SPARQL: needs preprocessing

SQL: use aggregate functions/GROUP BY
Granularity: Choosing the Right Level of Abstraction

1995 gab Edzard Reuter den Vorstandsvorsitz der Daimler Benz AG an Schrempp ab.

1995 ≤ t ≤ ?: <js,musing:ceo0f,db> – right??

what is meant by 1995, given this context?

✓ 1995-01-01(T00:00:00) – nope
✓ somewhere in 1995

• there exists an interval that starts in 1995 in which JS was CEO
• ceoship probably continues in 1996 → OpenRightInterval
The 1995 Example: Granularity, cont.

- find the right granularity
  - say, we are talking about things no finer than year, month, and day
- 1995 is translated into an instance of ProperInstantDay
- ProperInstantDay says that year, month, and day are functional properties (cardinality: 0 or 1)
- slot filler for year: 1995
- i.e., interpret this instant as an „underspecified“ existential constraint on the starting time of the interval, since month and day are not specified
More Granularity


- two instances \( b \) and \( e \) of ProperInstantDay
- 1995 is slot filler for year in \( b \), 2005 for year in \( e \)
- ClosedInterval \( i \) with
  - \( \text{begins}(i) = b \)
  - \( \text{ends}(i) = e \)
- further (textual) information might complete month and day of both \( b \) and \( e \) in \( i \)
Advantages

- properties that do not change over time can be relocated from `TimeSlice` to `Perdurant` (no duplication of information)
- the subtypes of `TimeSlice` (e.g., `Company`, `Person`, etc.) specify the behavior of a perdurant in a certain time interval (company, person, etc.)
- since `hasTimeSlice` is typed to `TimeSlice`, different slices need not to be of the same type
  - e.g., perdurant SRI has a time slice for `Company` and a slice for `AcademicInstitution`
  - i.e., a perdurant/entity can act in different ways
Advantages—Two Examples

- given time slices for a perdurant, we can infer useful (implicit) knowledge
  - two time slices $s, t$ for DaimlerChrysler
    - time interval $i$ of $s$ contains $j$ of $t$
    - $s$ specifies address for DC, $t$ does not
    - assume that subinterval inheritance holds for `hasAddress`
      - effect: address of DC at $j$ is equal to that of DC at $i$
  - two time slices $s, t$ for Jürgen Schrempp
    - both slices say that JS is CEO of DC
    - time interval $i$ of $s$ is strictly smaller than $j$ of $t$
    - $\exists k \ s.t. i \leq k \leq j$, where JS is very probably CEO of DC in $k$
Advantages, cont.

- higher-order properties/modalities
  - know, believe, ...
  - Ich glaube, dass Jürgen Schrempp zum 31. Dezember als Vorstandsvorsitzender von DC zurücktreten wird.
  - time slice p3 of perdurant i (ich) has property believe with time slice p2

JS resigns from DC – right semantics?
Finding the Right Semantics – Correction


No, JS resigns from DC´s ceoship!

\[\text{js} \xrightarrow{\text{resignsFrom}} \text{p1} \xrightarrow{\text{ceoOf}} \text{c1} \]

\[\text{hasTimeSlice} \]

\[\text{hasTemporalEntity} \]

\[\text{dc} \]

\[\text{oli1} \]

\[\langle __, 2005-12-31 \rangle \]

\[\text{hasTimeSlice} \]

\[\text{hasTemporalEntity} \]

\[\text{pid1} \]

\[2005-12-31 \]
A Unified Reasoning Architecture

Looking for Software Systems that Do the Right Thing
Different Kinds of Reasoning

- OWL
  - taxonomic axioms, weak property language
  - assertional knowledge
  - "built-in" TBox/ABox reasoning
- rule knowledge (local context)
  - more than two variables involved, numerical constraints, arithmetics
    - if X takes over position Y from Z at T
      then X has position Z from T on and Y had Z until T
    - if individuals X and Y have crucial properties in common
      then X sameAs Y
    - if X is a Person and X has annual income > 10,000,000 €
      then X is a VIP
“Reasoning” with Queries

- “global” knowledge involving many individuals
  - multiple overlapping intervals state that property P holds for X: combine into a single interval, using min and max
  - would like to see SQL-like aggregates & GROUP BY
  - might be done with rules, provided that functors are available
    - but: introduces large amounts of uninteresting facts and is therefore impractical
Requirements for Software

- what's needed:
  - triple store / OWL reasoner that scales up well
  - rule reasoning component
  - query component (preferably SPARQL)
- freely available systems only
- there's no single system which provides that, so:
- combine the most promising candidates
Finding a Compromise

- MUSING ontologies are justs about to be settled
- only small sets of preliminary test data
- use an available mid-size ontology instead
  - LT-World contains classes and facts about Language Technology areas, people, and institutions
  - 3,400 classes, 380 properties, 9,000 instances
  - ontology contents are the base of www.ltword.org
Candidate Systems

- OWLIM (v2.9.0 from www.ontotext.com)
  - has been (partly) developed in other EU projects, inference layer to Sesame (www.openrdf.org)
- Jena (v2.5.2, jena.sourceforge.net)
  - originally developed at HP, now open source
- Pellet (v1.5.0 pellet.owldl.com)
  - developed at Univ. of Maryland, now clarkparsia.com
- RacerPro (v1.9, test licence)
  - excluded because of memory overflow while loading test ontology
OWLIM

- by far the fastest triple store and OWL reasoner, when load and query times are taken into account
- rule compiler TRREE freely available but no source code
- restricted rule language, no functions or numerical constraints
- query language (at the moment) SeRQL (Sesame)
- pure forward reasoning (total materialization)
Jena

- OWL reasoning much slower than in OWLIM
- mostly forward reasoning, backward rules are also possible (tabling)
- rule language is more expressive
- SPARQL query language (almost standard)
- JenaSesameBridge allows to use Sesame (and OWLIM) as a model in Jena
Pellet

- description logic reasoner for OWL DL (OWL 1.1)
- tableaux-based reasoner
- very useful for consistency checks
  - instructive error messages
- already integrated with Jena
System Architecture

- all components are integrated as Jena models
- this allows to easily test and exchange components, even at runtime, if desired
- since the initial tests are artificial, the system can later be adapted to the real needs
- only OWL and rule inferencing tested
System Architecture, cont.

- **initial ontology** → **Pellet**
- **compile time**:
  - **fixpoint** → **Jena**
  - **QDP**
  - **result table**
- **run time**:
  - **XBRL BS** → **information extraction** → **NL text**
Initial Experimental Results

- OWLIM, Pellet and Jena OWL reasoners as base models
- Jena as rule inference model and query engine
- LT-World ontology and very small custom ruleset as test data
- best performance with
  - OWLIM as OWL reasoner and limited rule engine
  - Jena as Rule Inference Engine and Query Processor
Experimental Results, Numbers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OWLIM+Jena</td>
<td>49</td>
<td>115</td>
<td>0.27</td>
</tr>
<tr>
<td>Pellet+Jena</td>
<td>80</td>
<td>1,640</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Pentium 4, 2GHz, 1GB Ram
References

- spacetime worms, perdurant, time slice

- OWL-Time

- PROTON upper-level ontology
  - http://proton.semanticweb.org
Human Language Technology in Musing
Human Language Technology in Business Intelligence

- Business Intelligence (BI) is the process of finding, gathering, aggregating, and analysing information for decision making
  - Many systems in BI are portals which allow business analysts access to information
  - It is the work of the business analyst to dig into the documents in order to extract useful facts for decision making

- Analytical techniques traditionally used in BI rely on structured information and hardly ever use qualitative information which the industry is keen in using (e.g. opinions)

- It is important to make use of structured, semi-structured, and unstructured sources for decision making: because information is usually distributed across sources, it is unlikely that the sought after information will be found in one source

- Methods are required to make different sources interoperable for analysis
Proposed Solution

- Apply Human Language Technology to transform unstructured sources into the structured knowledge more suitable for analysis
- Content mining using domain-specific ontologies which precisely define the application domain
- Enables extraction of relevant information to be fed into models for financial risk analysis (credit rating, etc.), partner search for business, competitor monitoring, etc.
- Use ontology and standards for business reporting, for information exchange
Information Extraction (IE)

- IE pulls facts from the document collection
- It is based on the idea of scenario template
  - some domains can be represented in the form of one or more templates
  - templates contain slots representing semantic information
  - IE instantiates the slots with values: strings from the text or associated values
- IE is domain dependent – a template has to be defined
- Message Understanding Conferences 1987-1997 fuelled the IE field and made possible advances in techniques such as Named Entity Recognition
- From 2000 the Automatic Content Extraction (ACE) Programme
SENER and Abu Dhabi’s $15 billion renewable energy company MASDAR new joint venture Torresol Energy has announced an ambitious solar power initiative to develop, build and operate large Concentrated Solar Power (CSP) plants worldwide..... SENER Grupo de Ingeniería will control 60% of Torresol Energy and MASDAR, the remaining 40%. The Spanish holding will contribute all its experience in the design of high technology that has positioned it as a leader in world engineering. For its part, MASDAR will contribute with this initiative to diversifying Abu Dhabi’s economy and strengthening the country’s image as an active agent in the global fight for the sustainable development of the Planet.

<table>
<thead>
<tr>
<th>COMPANY-1</th>
<th>SENER</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPANY-2</td>
<td>MASDAR</td>
</tr>
<tr>
<td>% COMP-1</td>
<td>60%</td>
</tr>
<tr>
<td>% COMP-2</td>
<td>40%</td>
</tr>
<tr>
<td>NEW COMPANY</td>
<td>Torresol Energy</td>
</tr>
<tr>
<td>PURPOSE</td>
<td>“...develop, build, and operate CSP plants worldwide...”</td>
</tr>
</tbody>
</table>
Uses of the extracted information

- Template can be used to populate a data base (slots in the template mapped to the DB schema)
- Template can be used to generate a short summary of the input text
  - “SENER and MASDAR will form a joint venture to develop, build, and operate CSP plants”
- Data base can be used to perform querying/reasoning
  - Want all company agreements where company X is the principal investor
Information Extraction Tasks

- Named Entity recognition (NE)
  - Finds and classifies names in text
- Coreference Resolution (CO)
  - Identifies identity relations between entities in texts
- Template Element construction (TE)
  - Adds descriptive information to NE results
- Scenario Template production (ST)
  - Instantiate scenarios using TEs
Examples

- **NE:**
  - SENER, SENER Grupo de Ingenieria, Abu Dhabi, $15 billion, Torresol Energy, MASDAR, etc.

- **CO:**
  - SENER = SENER Grupo de Ingenieria = The Spanish holding

- **TE:**
  - SENER (based in Spain); MASDAR (based in Abu Dhabi), etc.

- **ST**
  - combine entities in one scenario (as shown in the example)
Named Entity Recognition

- It is the cornerstone of many NLP applications – in particular of IE
- Identification of named entities in text
- Classification of the found strings in categories or types
- General types are Person Names, Organizations, Locations
- Others are Dates, Numbers, e-mails, Addresses, etc.
- Domains may have specific NEs: film names, drug names, programming languages, names of proteins, etc.
Approaches to NER

- Two approaches:
  - (1) Knowledge-based based on humans defining rules;
  - (2) Machine learning approach, possibly using an annotated corpus

- Knowledge-based approach
  - Word level information is useful in recognising entities:
    - capitalization, type of word (number, symbol)
  - Specialized lexicons (Gazetteer lists) usually created by hand; although methods exist to compile them from corpora
    - List of known continents, countries, cities, person first names
    - On-line resources are available to pull out that information
Approaches to NER

- Knowledge-based approach
  - rules are used to combine different evidences
  - a known first name followed by a sequence of words with upper initial may indicate a person name
  - a upper initial word followed by a company designator (e.g., Co., Ltd.) may indicate a company name
  - a cascade approach is generally used where some basic names are first identified and are latter combined into more complex names
Approaches to NER

- **Machine Learning Approach**
  - Given a corpus annotated with named entities we want to create a classifier which decides if a string of text is a NE or not
    - ...<person>Mr. John Smith</person>...
    - ...<date>16th May 2005</date>
  - The problem of recognising NEs can be seen as a classification problem.
Machine Learning Approach

- Each named entity instance is transformed for the learning problem
  - ...<person>Mr. John Smith</person>...
  - Mr. is the beginning of the NE person
  - Smith is the end of the NE person

- The problem is transformed in a binary classification problem
  - is token begin of NE person?
  - is token end of NE person?

- The token itself and context are used as features for the classifier
Name Entity Recognition
Performance Evaluation

- Evaluation metric – mathematically defines how to measure the system’s performance against a human-annotated, gold standard.

- Scoring program – implements the metric and provides performance measures:
  - For each document and over the entire corpus
  - For each type of NE
The Evaluation Metric

- Precision = correct answers/answers produced
- Recall = correct answers/total possible correct answers
- Trade-off between precision and recall
- F-Measure = \((\beta^2 + 1)PR / \beta^2R + P\)
  
  [van Rijsbergen 75]
  
  - \(\beta\) reflects the weighting between precision and recall, typically \(\beta=1\)
Linguistic Processors in IE

- Tokenisation and sentence identification
- Parts-of-speech tagging
- Morphological analysis
- Name entity recognition
- Full or partial parsing and semantic interpretation
- Discourse analysis (co-reference resolution)
Approaches to information extraction

- **Extraction patterns**
  - “X announced a join venture agreement with Y”
  - “A joint venture between X and Y”
  - “The company will be called Z”

- **Hand-crafted systems**
  - Computational linguist writes rules based on corpus analysis and linguistic intuition

- **Machine Learning systems**
  - Learning a dictionary of information extraction patterns
  - Learning rules to tag start/end of semantic tags
  - Learning a tagging system using HMM
  - Applying statistical methods (SVM)
System development cycle

1. Define the extraction task
2. Collect representative corpus (set of documents)
3. Manually annotate the corpus to create a gold standard
4. Create system based on a part of the corpus: create identification and extraction rules
5. Evaluate performance against part of the gold standard
6. Return to step 3, until desired performance is reached
“Gold standard” corpora are divided typically into a training, sometimes testing, and unseen evaluation portion.

Rules and/or ML algorithms developed on the training part.

Tuned on the testing portion in order to optimise:

- Rule priorities, rules effectiveness, etc.
- Parameters of the learning algorithm and the features used.

Evaluation set – the best system configuration is run on this data and the system performance is obtained.

No further tuning once evaluation set is used!
GATE (Cunningham & al’02) General Architecture for Text Engineering

- Framework for development and deployment of natural language processing applications
  - http://gate.ac.uk
- A graphical user interface allows users (computational linguists) access, composition and visualisation of different components and experimentation
- A Java library (gate.jar) for programmers to implement and pack applications
Component Model

- **Language Resources (LR)**
  - data

- **Processing Resources (PR)**
  - algorithms

- **Visualisation Resources (VR)**
  - graphical user interfaces (GUI)

- Components are extendable and user-customisable
  - for example adaptation of an information extraction application to a new domain
  - to a new language where the change involves adaptation of a module for word recognition and sentence recognition
Documents in GATE

- A document is created from a file located somewhere in your disk or in a remote place or from a string.
- A GATE document contains the “text” of your file and sets of annotations.
- When the document is created and if a format analyser for your type is available “parsing” (format) will be applied and annotations will be created:
  - xml, sgml, html, etc.
- Documents also store features, useful for representing metadata about the document:
  - some features are created by GATE.
- GATE documents and annotations are LRs.
Documents in GATE

- Annotations have
  - types (e.g. Token)
  - belong to particular annotation sets
  - start and end offsets – where in the document
  - features and values which are used to store orthographic, grammatical, semantic information, etc.

- Documents can be grouped in a **Corpus**

- Corpus is other language resource in GATE which implements a set of documents
Documents in GATE

- semantics
- names in text
- information
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  <!-- XSchema definition for token-->
  <element name="Address">
    <complexType>
      <attribute name="kind" use="optional">
        <simpleType>
          <restriction base="string">
            <enumeration value="email"/>
            <enumeration value="url"/>
            <enumeration value="phone"/>
            <enumeration value="ip"/>
            <enumeration value="street"/>
            <enumeration value="postcode"/>
            <enumeration value="country"/>
            <enumeration value="complete"/>
          </restriction>
        </simpleType>
      </attribute>
    </complexType>
  </element>
</schema>
Manual Annotation in GATE GUI
Annotation in GATE GUI

The following tasks can be carried out manually in the GATE GUI:

☑ Adding annotation sets
☑ Adding annotations
☑ Resizing them (changing boundaries)
☑ Deleting
☑ Changing highlighting colour
☑ Setting features and their values
Preserving and exporting results

- Annotations can be stored as stand-off markup or in-line annotations.
- The default method is standoff markup, where the annotations are stored separately from the text, so that the original text is not modified.
- A corpus can also be saved as a regular or searchable (indexed) datastore.
Text Processing Tools in GATE

- Document Structure Analysis
  - Different document parsers take care of the structure of your document (xml, html, etc.)
- Tokenisation
- Sentence Identification
- Parts of speech tagging
- (many more processors)

- All these resources have as runtime parameter a GATE document, and they will produce annotations over it
- Most resources have initialisation parameters
Rule-based NE recognition in GATE

☑ In GATE Gazetteers lists entries may contain some useful semantic information
  • for example one may associate some features and values to entry names
  • features can be used in grammars or can be used to enrich system output
  • gazetteer lists are organized in index files
Named Entity Grammar in GATE

- Implemented in the JAPE language (part of GATE)
  - Regular expressions over annotations
  - Provide access and manipulation of annotations produced by other modules
- Rules are stored in grammar files
- Grammar files are compiled into Finite State Machines
- A main grammar files specifies how different grammars should be executed (phases)
  - constitute a cascade of FSTs over annotations
NER in GATE

- Rules are **hand-coded**, so some linguistic expertise is needed here
- uses annotations from tokeniser, POS tagger, and gazetteer modules
- use of contextual information
- rule priority based on pattern length, rule status and rule ordering
- Common entities: persons, locations, organisations, dates, addresses.
JAPE Language

- A JAPE grammar rule consists of a left hand side (LHS) and a right hand side (RHS)
  - LHS = what to match (the pattern)
  - RHS = how to annotate the found sequence
  - LHS -> RHS

- A JAPE grammar is a sequence of grammar rules
- Grammars are compiled into finite state machines
- Rules have priority (number)
- There is a way to control how to match
  - options parameter in the grammar files
JAPE Grammar

- In a file with name something.jape we write a Jape grammar (phase)

Phase: example1
Input: Token Lookup
Options: control = appelt

Rule: PersonMale
Priority: 10
(  
{Lookup.majorType == first_name, Lookup.minorType == male}
({Token.orth == upperInitial})*)
):annotate
-->
:annotate.Person = { gender = male }

....(more rules here)
Main JAPE grammar

- Combines a number of single JAPE files in general named “main.jape”

  MultiPhase: CascadeOfGrammars
  Phases:
  grammar1
  grammar2
  grammar3
ANNIE System

- A Nearly New Information Extraction System
  - recognizes named entities in text
  - "packed" application combining/sequencing the following components: document reset, tokeniser, splitter, tagger, gazetteer lookup, NE grammars, name coreference
  - can be used as starting point to develop a new name entity recogniser
Semantic Annotation: Motivation

- Semantic metadata extraction and annotation is the glue that ties ontologies into document spaces
- Metadata is the link between knowledge and its management
- Manual metadata production cost is too high
- State-of-the-art in automatic annotation needs extending to target ontologies and scale to industrial document stores and the web
Once metadata is attached to documents, they become much more useful and more easily processable, e.g. for categorising, finding relevant information, and monitoring.

Such metadata can be divided into two types of information: explicit and implicit.

Explicit metadata extraction involves information describing the document, such as that contained in the header information of HTML documents (titles, abstracts, authors, creation date, etc.)

Implicit metadata extraction involves semantic information deduced from the text, i.e. endogenous information such as names of entities and relations contained in the text. This essentially involves Information Extraction techniques, often with the help of an ontology.
Metadata extraction (2)

- a hierarchy added to the set of semantic tags;
- a hierarchy of relations;
- there are usually more tags than before!
- there are inference mechanisms in the background;
- there is a knowledge base of known facts, e.g.:

  "London" <capital-of> "UK" <located-in> "Western Europe" <part-of> "Europe"

- new searches possible: “Companies located in Western Europe?”
Ontology Learning and Population: Motivation

- Creating and populating ontologies manually is a very time-consuming and labour-intensive task.
- It requires both domain and ontology experts.
- Manually created ontologies are generally not compatible with other ontologies, so reduce interoperability and reuse.
- Manual methods are impossible with very large amounts of data.
Semantic Annotation vs Ontology Population

- **Semantic Annotation**
  - Mentions of instances in the text are annotated wrt concepts (classes) in the ontology.
  - Requires that instances are disambiguated.
  - It is the text which is modified.

- **Ontology Population**
  - Generates new instances in an ontology from a text.
  - Links unique mentions of instances in the text to instances of concepts in the ontology.
  - It is the ontology which is modified.
Ontology-based Information Extraction (OBIE)

- Traditional IE is based on a flat structure, e.g. recognising Person, Location, Organisation, Date, Time etc.

- For semantic-based richer access to information, we need information in a hierarchical structure

- Idea is that we attach semantic metadata to the documents, pointing to concepts in an ontology

- Information can be exported as an ontology annotated with instances, or as text annotated with links to the ontology
MUSING applications requiring HLT

- A number of applications have been specified to demonstrate the use of semantic-based technology in BI – some examples include
  - Collecting Company Information from multiple multilingual sources (English, German, Italian) to provide up-to-date information on competitors
  - Identifying Chances of success in regions in a particular country
  - Semi-automatic form filling in serveral Musing applications
  - Identify appropriate partners to do business with
  - Creation of a Joint Ventures Database from multiple sources
Natural Language Processing Technology

- Main components adapted for MUSING applications are gazetteer lists and grammars used for named entity recognition.
- New components include:
  - An ontology mapping component – entities are mapped into specific classes in the given ontology.
  - A component creates RDF statements for ontology population based on the application specification:
    - For example, create a company instance with all its properties as found in the text.
Ontology-based IE in MUSING

DATA SOURCE PROVIDER -> DOCUMENT

DOCUMENT COLLECTOR -> DOCUMENT

MUSING DATA REPOSITORY -> MANUALLY ANNOTATED DOCUMENTS

MANUALLY ANNOTATED DOCUMENTS -> ANNOTATION TOOL

ANNOTATION TOOL -> DOMAIN EXPERT

DOMAIN EXPERT -> MUSING ONTOLOGY CURATOR

MUSING ONTOLOGY CURATOR -> USER

USER INPUT -> USER

MUSING APPLICATION

ECONOMIC INDICATORS -> REGION SELECTION MODEL

REGION SELECTION MODEL -> REGION RANK

REGION RANK -> COMPANY INFORMATION

COMPANY INFORMATION -> ENTERPRISE INTELLIGENCE

ENTERPRISE INTELLIGENCE -> REPORT

REPORT -> KNOWLEDGE BASE

KNOWLEDGE BASE -> INSTANCES & RELATIONS

MANUALLY ANNOTATED DOCUMENTS -> ONTOLOGY POPULATION

ONTOLOGY POPULATION -> ANNOTATED DOCUMENT

ANNOTATED DOCUMENT -> ONTOLOGY-BASED INFORMATION EXTRACTION SYSTEM

ONTOLOGY-BASED INFORMATION EXTRACTION SYSTEM -> DOCUMENT

DOCUMENT -> MUSING ONTOLOGY

MUSING ONTOLOGY -> USER

USER INPUT -> USER
Company Information in MUSING
Extracting Company Information

- Extracting information about a company requires for example identifying the Company Name; Company Address; Parent Organization; Shareholders; etc.
- These associated pieces of information should be asserted as properties values of the company instance.
- Statements for populating the ontology need to be created ("Alcoa Inc" hasAlias "Alcoa"; "Alcoa Inc" hasWebPage "http://www.alcoa.com", etc.)
Region Selection Application

- Given information on a company and the desired form of internationalisation (e.g., export, direct investment, alliance) the application provides a ranking of regions which indicate the most suitable places for the type of business.

- A number of social, political geographical and economic indicators or variables such as the surface, labour costs, tax rates, population, literacy rates, etc. of regions have to be collected to feed an statistical model.
Region Information

- Indicators such as:
  - Economic Stability Indicators: exports, imports, etc.
  - Industry Indicators: presence of foreign firms, number of procedures to start business, etc.
  - Infrastructure Indicators: drinking water, length of highway system, hospitals, telephones, etc.
  - Labour Availability Indicators: employment rate, libraries, medical colleges,
  - Market Size Indicators: GDP, surface, etc.
  - Resources Indicator: Agricultural land, Forest, number of strikes, etc.
Region Information - examples

- “the net irrigated area totals 33,500 square kilometres” and “The land drained by these rivers is agriculturally rich” – AGRIC-LAND (agricultural land)
- “Males constitute 50.3 million” – URBM (urban population)
- “64.14% of the people are employed and allied activities” – EMP (employment)
- “The three airports in Himachal Pradesh are…..” – AIRP_V (air freight)
- “In rural areas over 65% of the population have no access to safe drinking water” – WCHAN (water challenges)
Region Selection Application

- Data sources used for the OBIE application are statistics from governmental sources and available region profiles found on the Web (e.g. Wikipedia)
- Gazetteer lists contain location names and associated information together with keywords to help identify the key information
- Grammars use contextual information and named entities to identify the target variables
  - “unemployment rate of 25% (2001)”
- Extraction performance obtained: F-score > 80%
Extracting Economic Indicators
Walk-through Example

From the Wikipedia article on *Andhra Pradesh* (a province of India):

- Andhra Pradesh has 1330 Arts, Science and Commerce colleges, 238 Engineering colleges and 53 Medical colleges. The student to teacher ratio is 19:1 in the higher education. According to census taken in 2001, Andhra Pradesh has an overall literacy rate of 60.5%. While male literacy rate is at 70.3%, the female literacy rate however is only at 50.4%, a cause for concern.
According to census taken in 2001, Andhra Pradesh has an overall literacy rate of 60.5%.
According to census taken in 2001, Andhra Pradesh has an overall literacy rate of 60.5%.

- **Type**: Mention
- **Features**
  - `article_region_code`: India_AP
  - `indicator_value`: 60.50%
  - `key`: LIT_T
  - `year`: 2001
Example

According to census taken in 2001, Andhra Pradesh has an overall literacy rate of 60.5%.

<table>
<thead>
<tr>
<th>Type</th>
<th>Mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>article_region_code</td>
<td>India_AP</td>
</tr>
<tr>
<td>region_instance</td>
<td><a href="http://musing.deri.at/ontologies/v0.5/int/region#AndhraPradesh">http://musing.deri.at/ontologies/v0.5/int/region#AndhraPradesh</a></td>
</tr>
<tr>
<td>indicator_value</td>
<td>60.50%</td>
</tr>
<tr>
<td>key</td>
<td>LIT_T</td>
</tr>
<tr>
<td>indicator_instance</td>
<td><a href="http://musing.deri.at/ontologies/v0.5/int/indicator#LIT_T">http://musing.deri.at/ontologies/v0.5/int/indicator#LIT_T</a></td>
</tr>
<tr>
<td>year</td>
<td>2001</td>
</tr>
</tbody>
</table>
RDF output

A custom PR checks the features of the Mention annotation and fills in an appropriate template to generate RDF.

This RDF will create an instance of Measurement with appropriate property values, so the knowledge base can be updated with the extracted information.
RDF output

```xml
<indicator:Measurement rdf:ID="Measurement_173">
  <time:hasTimeSlice>
    <time:TimeSlice rdf:ID="TimeSlice_91">
      <time:hasTemporalEntity>
        <time:ProperInstantYear rdf:ID="ProperInstantYear_33">
          <time:hasTimeSlice>
            <time:year rdf:datatype="http://www.w3.org/2001/XMLSchema#int">2001</time:year>
          </time:hasTimeSlice>
        </time:ProperInstantYear>
        <time:hasTemporalEntity>
          <time:TimeSlice>
            <indicator:hasValue rdf:datatype="http://www.w3.org/2001/XMLSchema#string">60.5%</indicator:hasValue>
            <indicator:hasPoliticalRegion rdf:resource="http://musing.deri.at/ontologies/v0.5/int/region#AndhraPradesh"/>
            <indicator:hasIndicator rdf:resource="http://musing.deri.at/ontologies/v0.5/int/indicator#LIT_T"/>
          </time:TimeSlice>
        </time:hasTemporalEntity>
      </time:hasTemporalEntity>
    </time:TimeSlice>
  </time:hasTimeSlice>
</indicator:Measurement>
```
Creation of Gold Standards with an Annotation Tool

- Web-based Tool for Ontology-based (Human) Annotation
  - User can select a document from a pool of documents
  - Load an ontology
  - Annotate pieces of text wrt ontology
  - Correct/save the results back to the pool of documents
Joint Venture Annotation

AS20040924013CAA
24/09/2004
AGRISOLE
Venerdì
PRIMA PAGINA
Merlini, joint bulgara per i funghi biologici
Franco Rufolo
VERONA - Anche il fungo spontaneo diventa biologico. Una certificazione italiana, infatti, identifica un'area precisa - in Bulgaria, in joint venture con la veronese Riccardo Merlini - dove, proprio per le caratteristiche ambientali (niente trattamenti da almeno tre anni e una raccolta che non comprometta l'habitat naturale), il prodotto micologico ha tutte le caratteristiche bio. Nascono così i funghi porcini di bosco da agricoltura biologica. E' una nuova frontiera di qualità' di cui si interessa l'industria guidata da Davide Merlini, affiancato dai fratelli Diego e Laura, mentre il fondatore Riccardo continua nella sua passione di 'raccolgere' di porcini, ovuli e spugnole, e segue le orme del fondatore. <BR>
Primi funghi bio, quindi, a conclusione di un'annata, quella del 2003, scarsa (così' la definisce il direttore commerciale Diego Merlini) a causa della siccità' e altre componenti che ne hanno compromesso la raccolta in tutta Europa. La campagna in corso, invece, si presenta all'insegna della buona qualità'. Ricorda ancora Diego Merlini: '&quot;ci avvaliamo delle attivita' di due societá' in joint venture con noi - in Bulgaria e in Serbinia' - il cui lavoro per noi e' fondamentale. Trovandosi in zone di raccolta, infatti, ci consentono di monitorare al meglio il mercato. Il contatto diretto con i luoghi di raccolta e' indispensabile per un prodotto come il nostro'&quot; <BR>
Malgrado il quadro produttivo generale, la Merlini si e' confermata primo produttore mondiale nella produzione di funghi sportanei, con la raccolta e la lavorazione di 70.000 quintali di prodotto fresco,
AS20040924013CAA
24/09/2004
AGRISOLE
Venerdì
PRIMA PAGINA
Merlini, joint bulgara per i funghi biologici
Franco Ruffo
VERONA - Anche il fungo spontaneo diventa biologico. Una certificazione italiana, infatti, identifica un'area precisa - in Bulgaria, in joint venture con la veronese Riccardo Merlini - dove, proprio per le caratteristiche ambientali e l'habitat naturale, il bosco da agricoltura Davide Merlini, affilia passione di &quot;green&quot; e direttore commerciale del gruppo, ha avviato un'operazione di riciclo. Buona qualità! Riccardo Merlini, con noi - in joint venture - raccolta, infatti, ci consente di produrre in tonnellate di funghi, indispensabili a Iain G. Smith, direttore scientifico dell'Istituto di Farmacologia e Biologia della Natura.
Malgrado il quadro produttivo generale, la Merlini si è confermata primo produttore mondiale nella produzione di funghi spontanei, con la raccolta e la lavorazione di 70.000 quintali di prodotto fresco,
Region Information Annotation

GATE 4.0 build 2823

Annotation Sets  Annotations List  Co-reference Editor  OAT  Text

Himachal Pradesh - Wikipedia, the free encyclopedia

Himachal Pradesh
From Wikipedia, the free encyclopedia
Jump to: navigation, search

Himachal Pradesh
Capital
- Coordinates Shimla
- 30.06° N 77.11° E
Largest city Shimla
Population (2001)
- Density 1,097,248 (20th)

Area
- 24,612 sq km
- 12

POP
- Establishment
- Government
- Chief Minister
- Legislature
- V. S. Kochhar
- Virbhadra Singh
- Unicameral
- Official Language

ontology: http://gate.ac.uk/owlim
inst: http://www.owl-ontologies.com/Ontology1182768946.owl#POP
Due to the abundance of perennial rivers, Himachal also sells hydro electricity to other states such as Delhi, Punjab & Rajasthan.

Himachal is being mooted as a possible host for the 2010 Commonwealth Winter Games.

The geography of Himachal presents considerable challenge to the development of transport infrastructure. Nevertheless, the state has made significant progress in road connectivity in the last few decades. Himachal at present has the highest road density among all the hill states of India. Although Himachal also has three airports and two narrow gauge rail tracks, roads remain the main mode of transport in Himachal.

Eight national highways (NH) pass through the state with a total length of 1235 km. NH 1-A touches Shahpur. NH 1 passes through Kangra. NH 2A passes through Mandi. NH 3 passes through Rampur. NH 4 goes through Chamba. NH 5 passes through Kullu. NH 6 passes through Una. NH 7 passes through Mahipalpur. NH 8 passes through Pathankot. NH 9 passes through Thal. NH 10 passes through Solan.
Tools to develop the extraction system

- Given a set of documents (corpus) human-annotated, we can index the documents using the human and automatic annotations (e.g. tokens, lookups, pos) with the ANNIC tool.
- The developer can then devise semantic tagging rules by observing annotations in context.
- Another alternative is to use ML capabilities of the GATE system – supervised learning.
Identity Resolution in MUSING

- Same Person Name different Entity

- P1) **Antony John** was born in 1960 in Gilfach Goch, a mining town in the Rhondda Valley in Wales. He moved to Canada in 1970 where the woodlands and seasons of Southwestern Ontario provided a new experience for the young naturalist...

- P2) **Antony John** - Managing Director. After working for National Westminster Bank for six years, in 1986, Antony established a private financial service practice. For 10 years he worked as a Director of Hill Samuel Asset Management and between 1999 and 2003 he was an Executive Director at the private Swiss bank, Lombard Odier Darier Hentsch. Antony joined IMS in 2003 as a Partner. Antony’s PA is Heidi Beasley...
Identity Resolution in MUSING

- Same company name, different company

- C1) Operating in the market where knowledge processes meet software development, **Metaware** can support organizations in their attempts to become more competitive. Metaware combines its knowledge of company processes and information technology in its services and software. By using intranet and workflow applications, Metaware offers solutions for quality control, document management, knowledge management, complaints management, and continuous improvement.

- C2) **Metaware S.r.l.** is a small but highly technical software house specialized in engineering software and systems solutions based on internet and distributed systems technology. Metaware has participated in a number of RTD cooperative projects and has a consolidated partnership relationship with Engineering.
Approaches to Identity Resolution in MUSING

- Text based approach
  - clustering informed by semantic analysis and summarization
  - extract sentences containing entity of interest and create a summary
  - extract semantic information from summaries and create term vectors for clustering
  - apply agglomerative clustering to the set of vectors
  - good performance on Person information
Approaches to Identity Resolution in MUSING

- Ontology-based approach
  - define rules for each class in the ontology
  - rules combine different similarity criteria using a weighting mechanism:
    - compare alias name ("Alcoa" vs "Alcoa Inc.")
    - compare location (Scotland is in the UK)
    - etc.
  - select candidate instances from ontology
  - compare target instance to each candidate
  - evaluation of merging information extracted from company profiles;
    - performance ~ 89% (measure correctness of match)
Opinion Mining in Business Intelligence

- Opinion mining (OM) consists on identifying what opinion a particular discourse expresses (it is not interested with what the text is about).

- In MUSING we are interested in tracking opinions about business entities: persons, organizations, products & services, etc.

- The field of OM is relatively new, however it is very active thanks to initiatives such as:
  - the TREC 2006 Blog mining for opinion retrieval
  - NTCIR Workshop on Evaluation of Information Access Technologies
  - Text Analysis Conference with an opinion summarization task
I would like to recommend TRUPRINT (http://www.truprint.co.uk/)
Photo printing service. I uploaded a set of 27 pictures at 10am on Friday and they were on the doorstep at 8.30 the next morning with perfect quality and very well packaged and I was able to track their progress on the website. This is not the first time I have used them and I have always been impressed, but they really excelled themselves this time. If you use a digital camera this is a very good service and only 10p a print.

positive opinions

Keep well away from Solar Technik.

My husband and I were approached by Solar Technik to install a solar system in our house. Having owned a Grade 2 Listed building, the rep promised us that if planning permission is not approved by the NCDC, our deposit will be refunded.

To cut to the chase, planning permission was refused. It's been almost 3 months now and we have not got our refund back. We made numerous calls to Solar Technik. The staff was discourteous and unhelpful, in particular the reception lady who apparently goes by 3 different names!!! We were also given all sorts of lame excuses.

In summing up, Solar Technik is simply an unprofessional company. Customer service is diabolical!! Never let them pressure you into buying their expensive system!

negative opinions

Date Reviewed: 26/12/07

michaelfrankl1
Member Since: 10/12/07

View Member's: Reviews

Delivery charge was not obvious until far too late. First attempt to enter order prompted error message, but did not tell me why the error had occurred and deleted address data already entered. This needs fast improvement!

negative opinion, but less evident
OM Approach

- OM can be approached as a classification problem
- Interested in
  - differentiate between positive opinion vs negative opinion
  - recognising fine grained evaluative texts (1-star to 5-star classification)
- We use a supervised learning approach (Support Vector Machines) that uses linguistic features
- The linguistic tools used in the Musing project apply various processors to produce words, roots, parts-of-speech, etc.
  - binary classification ~ 80% classification accuracy
  - fine grained classification ~ 74% classification accuracy
  - we don’t use specific ‘sentiment’ word list, the classifier learns from generic linguistic features