Natural Language Technology for Business Intelligence
Horacio Saggion & Adam Funk
Human Language Technology in Business Intelligence

• Business Intelligence (BI) is the process of finding, gathering, aggregating, and analysing information for decision making.

• BI has relied on structured/quantitative information for decision making and hardly ever use qualitative information found in unstructured sources which the industry is keen in using.

• NLP and BI:
  – **gathering** information through Information Extraction and Text analysis.
  – **aggregating** information through cross-source coreference, identity resolution, text clustering.

• This presentation will be based on part on our work for the MUSING EU project.
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 and has to be adapted to each application domain either manually or by machine learning
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 Grupo de Ingeniería</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>AGREEMENT</td>
<td>Joint Venture</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
Ontology-based Information Extraction

- The application domain (concepts, relations, instances, etc.) is modelled through an ontology or set of ontologies.
- Onto-based Information Extraction identifies in text instances of concepts and relations expressed in the ontology:
  - the extraction task is modelled through “RDF templates”
  - X is a company; Z is a person; Z is manager of X; etc.
  - It generally uses the ontology as input and output.
- Extracted information is used to populate a knowledge repository.
- Updating the KR involves a process of identity resolution.
- GATE components are particularly well adapted for Ontology-based IE:
  - in particular GATE has an API to manipulate the ontology and the ontology can be manipulated in extraction grammars.
Ontology-based IE in MUSING

DATA SOURCE PROVIDER

DOCUMENT COLLECTOR

MUSING DATA REPOSITORY

MANUALLY ANNOTATED DOCUMENTS

ANNOTATION TOOL

DOCUMENT

DOMAIN EXPERT

ONTODLOGY CURATOR

MUSING ONTOLOGY

USER

USER INPUT

MUSING APPLICATION

ECONOMIC INDICATORS

REGION SELECTION MODEL

REGION RANK

COMPANY INFORMATION

ENTERPRISE INTELLIGENCE

REPORT

MANUALLY ANNOTATED DOCUMENTS

ANNOTATION TOOL

ONTODLOGY POPULATION

ONTODLOGY BASE

INSTANCES & RELATIONS
Company Information in MUSING
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 – Italian financial news paper
  - Some English press data – Financial Times
  - Companies' web pages (main, “about us”, “contact us”, etc.)
  - Wikipedia, CIA Fact Book, etc.
  - CreditReform (data provider): company profiles; payment information – data provider
  - European Business Registry (data provider): profiles, appointments
  - Discussion forums
  - Log files for IT related applications

- Ontology is manually developed through interaction with domain experts and ontology curators
  - It extends the PROTON ontology and covers the financial, international, and IT operational risk domain
  - Particular methodology used to pull out the information from domain experts: “Competence Questions”
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
Joint Venture Annotation
Region Information Annotation

![GATE interface with region information annotation]

- **Himachal Pradesh** - Wikipedia, the free encyclopedia
  - Capital: Shimla
  - Coordinates: 30.91°N 77.11°E
  - Largest city: Shimla
  - Population: 3.2 million (2001)
  - Area: 22,671 km²

- **Time zone**: IST

- **Establishment**:
  - **Ontology**: http://gate.ac.uk/owlim
  - **Instance**: http://www.w3c-ontologies.com/ontologies/1182768845 owl#POP
Region Information Annotation

Due to the abundance of perennial rivers, Himachal also sells hydroelectricity to other states such as Delhi, Punjab, and 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 1236 km. NH-1A touches Shahrpaur. NH-512 goes outside Mandi.
Musing Applications

- Balance sheet conversion to XBRL for upload and navigation
- Credit rating for companies
- Obtain up-to-date information on competitors
- Monitor company agreements
- Estimate the chances of success of internationalisation ventures
- Compute the “reputation” of a business entity
- Most applications require semi-automatic fill in a questionnaire
  - Company main activity; Business Sector; Company code; Company foundation date; Products/Services; Partnerships; International Activities (import, export, etc.); Employee distribution (managers, office-workers, totals…); etc.
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 (e.g., "Alcoa Inc" hasAlias "Alcoa", "Alcoa Inc" hasWebPage "http://www.alcoa.com", etc.)
OBIE in Action

– Extracting Company Information DEMO
Some details

- Rule-based system
  - reuse of some default components for NE recognition + implementation of document structure analysers for each target source
  - lexicon/gazetteer list developed specifically for the application to identify keywords that mark presence of concepts
  - regular grammars that represent “typical” ways in which information (concepts, relations) is expressed in text
  - Mapping to ontology + RDF statements for Ontology population

- Current performance
  - F-score between ~ 80%
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, etc.
  - Market Size Indicators: GDP, surface, etc.
  - Resources Indicator: Agricultural land, Forest, number of strikes, etc.
Extracting Economic Indicators

- Extracting economic indicators from text
  DEMO
Extracting Economic Indicators

- The application uses the following components
  - Ontology-based gazetteer lookup process
  - Named entity recognition
  - Text classification based on SVM
  - Mention enrichment (features are added to each indicator)
  - RDF creation from identified information
  - Upload to the Knowledge Repository
Cross-source Coreference

- 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...
Cross-source Coreference

• 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.
SemEval 2007 Web People Search

- A search engine user types in a person name as a query
- Instead of ranking the Web pages, an ideal system should organize the results in as many clusters as different individuals sharing the name have been returned
- Input: Set of documents matching a person name
- Output: Clusters, each cluster refers to the same individual
- System participants have to carry out the task for a number of unseen names
- System output is compared to gold-standard data
SemEval Data/Metrics

- Metrics used are purity, inverse-purity, and the harmonic mean (f-score) of purity and inverse-purity
- The evaluation produced a ranking of systems where the systems are ranked according to the average f-score
- Training Data
  - for each person name there is a set of 100 documents
  - 10 person names from the European Conference on Digital Libraries
  - 7 person names from Wikipedia
  - 32 sets from a previous study (Gideon&Mann’03)
- Test Data
  - 30 people; pages found querying Yahoo! with the name of the person
Cross-source Coreference

- Text based approach: document analysis
  - input: set of documents and target name
  - agglomerative clustering algorithm which uses word-based and semantic-based “document representations”
  - process each document with the ANNIE system (GATE) and identify in documents names of type person, organization, address, date, location
  - in each document create coreference chains (only name coreference)
Cross-source Coreference

- Text based approach: representation
  - identify chains containing target entity to be disambiguated and extract sentences containing references to the target entity (~ summaries)
  - documents represented as vector of terms
    - terms: words or names
    - extracted from: full document or document summaries
  - use tf*idf weighting scheme
    - local idf tables for words and names created per set of documents
Cross-source Coreference

• Text based approach: clustering
  – initially as many clusters as documents
  – at each iteration the two more similar clusters are merged; if similarity < threshold stop the algorithm
  – similarity between two clusters ~ similarity between the two most similar docs in the cluster
    • we don’t create “centroids” but use the vectors in the cluster for comparison
  – use cosine similarity between vectors to decide if two documents are ‘similar’ (and therefore are “talking” about the same entity)
  – threshold estimated on 10 sets of training data
Cross-source Coreference

- Text based approach: evaluation
  - performance measured using purity, inverse purity, and f-score on SemEval test data
  - words + full documents = 0.74 f-score
  - words + summaries = 0.74 f-score
  - names + full documents = 0.68 f-score
  - names + summaries = 0.64 f-score
Cross-source Coreference

- Text based approach: re-evaluation
  - re-run algorithm considering the contribution of each name type
  - organization names + full documents = 0.78 f-score
  - person names + summaries = 0.70 f-score
  - improvement in both conditions (full document or summary) considering specific types of terms
Vector Computation Demo

• Clustering in Musing (IT-OpR domain) is been used to discover conceptual information by grouping related textual information from log-files. In this case vector computation is carried out over documents in Italian and a Web service is used to produce the analysis of Italian texts before they are used by our tools.
Ontology-based Identity Resolution

- Identity Resolution Framework using Ontology – Milena Yankova (OntoText)
  - input = entity + property values as specified in an ontology
  - output = updated ontology
  - identity rules are defined for each entity type in the ontology (e.g. companies, people)
  - rules combine different similarity criteria to compute a numeric score
Ontology-based Identity Resolution

- **Identity Resolution Framework**
  - **pre-filtering component**: select candidates from the ontology using some extracted properties found in text
    - for companies select those with some name similarity
  - **evidence collection component**: computes different identity criteria and produces a score
    - compute the distance between the company names
    - identify if one location (Scotland) is part of another location (UK)
  - **decision maker component**: decides on the most similar candidate
    - a similarity threshold is set optimising over training data (set at 0.40 for company information)
  - **data integration component**: updates the ontology
Ontology-based Identity Resolution

• Identity Resolution Experiments
  – ontology pre-populated with data from provider (database to ontology KB) – UK companies
  – UK company profiles feed to our company profile analyser to produce RDF templates for UK companies
  – Match attempted between extracted companies and the KB
    • $f$-score = 0.89