ONTORULE

Ontologiesmeet Business Rules

Integrating Ontologies and Rules in a Semantic Business: From Policies to Operation

Adeline Nazarenko Luis Polo Thomas Eiter Jos de Bruijn Antonia Schwichtenberg Stijn Heymans
(Paris 13, CTIC, TU Vienna, ontoprise)

30 May 2010 — ESWC 2010 Tutorials

Aim of this tutorial

You will understand:

1. How NLP can help to extract business knowledge from policy documents
2. What type of business knowledge is typically modeled as an ontology and what knowledge as rules (both logical and production rules)
3. Which approaches to combining ontologies and rules are currently available; which ones are implemented
4. How to choose the appropriate combination paradigm for your goals;
5. How to model a combination of ontologies and rules using mature commercial tools

Contents of the tutorial

1. From Business Policy Documents to a Business Model (1 hour 15 minutes)
2. Integrating Ontologies and Rules (2 hours 45 minutes)
   - 2.1 OWL 2 ontologies (45 minutes)
   - 2.2 Logic Programming and Production Rules (45 minutes)
   - 2.3 Integration (1 hour)
   - 2.4 Demo (15 minutes)

Acknowledgements

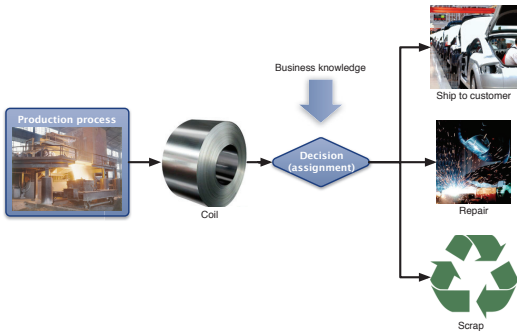
- The ONTORULE project consortium: ILOG (an IBM Company), ontoprise, Free University of Bolzano, Vienna University of Technology, PNA, Université Paris 13, Fundación CTIC, Audi and ArcelorMittal

  `http://ontorule-project.eu`

- This tutorial is co-funded by the European Union (FP7)

Business Vocabulary

Business Scenario



Part I

From Business Policy Documents to a Business Model

Overview of Part I

- Building business models from written policies
- State of the art
  - Ontology design
  - Ontology population
  - Information extraction
- Corpus design
- Ontology acquisition
  - Terminological analysis
  - Term normalizationion
  - Conceptualizationion
  - Relation identification
- Business rule acquisition
- Semantic Business Vocabulary and Rules

ONTORULE

Why starting from texts?

An alternative source of information

- Expert knowledge is difficult to make explicit
- Experts are seldom available for intensive interviews
- Large amount of written information is available
- Analysts usually exploit texts to prepare interviews

A challenge for knowledge management

- Business models must be documented
- Traceability of business models is often critical
- The source documents must be maintained together with the business models

ONTORULE

What can be extracted from texts?

Although

- Documents do not contain all the relevant information
- Domain knowledge is seldom explicit

Texts are precious source of information

- Documents often express factual knowlegde about the domain entities, their properties, their evolution and their relations.
- Documents also reflect the underlying domain knowledge : what are the relevant concepts? how do they relate to each other?
- Policy documents express the business rules that are relevant for the domain.

ONTORULE

Example of text

Product Definition

Major metallurgical families and characterization

ArcelorMittal's range of steels for the automotive sector comprises all the main metallurgical families:
- Steels for drawing: aluminum killed and IF (Interstitial Free);
- High- strength steels: high yield strength steels, rephosphorized steels, high strength IF steels, isotropic and bake hardening steels;
- Very high strength multiphase steels: Dual Phase, TRIP, ferrite- bainite, Multiphase steels.

The mechanical properties of these steels are the result of a combination of parameters that are defined throughout the steel manufacturing process. The two main parameters are:
- Chemical composition;
- Thermo- mechanical process.

Mechanical properties

To obtain the required mechanical properties, the steelmaker devises a range of strength/ formability combinations suitable for the uses to which products are to be put in the automobile.

A number of hardening processes are available. They can be employed alone or in combination:

Domain concept? Domain entity? Business rule? Domaine relation?

ONTORULE

Building a hierarchy of concepts

Two main approaches

- A distributional approach for "learning" ontologies
- A terminological approach for assisting the conceptualizationion task

ONTORULE

Distributional approach

The context in which the words occur reflect their meaning. Words with similar meanings tend to appear in similar contexts [Harris *et al.*, 1989]. Semantic classes of words can be built out of a distributional analysis.

Semantic classes extracted from texts are supposed to represent domain concepts but results are noisy and difficult to exploit.

[Hindle, 1990][Dagan *et al.*, 1993][Habert *et al.*, 1996][Nazarenko *et al.*, 1997][Faure and Nédellec, 1999][Maedche and Staab, 2001][Cimiano and Völker, 2005]

Example

mechanical properties
mechanical strength
mechanical process
mechanical behaviour

ONTORULE

Terminological approach

The terms are words and compounds form the domain vocabulary

Example

mechanical properties *coil* *yield point elongation* *strengh*

The terms extracted from a text reflect the domain vocabulary but a lot of manual work is necessary for filtering, structuring, modeling.

[Aussenac-Gilles *et al.*, 2000] [Aussenac-Gilles *et al.*, 2008] [Szulman *et al.*, 2009a]

ONTORULE

Ontology population

Given a conceptual model (ontological T-Box), named entity recognition tools can be used to extract the concept and relation instances and thus enrich the ontological A-Box.

[Magnini *et al.*, 2006] [Giuliano and Gliozzo, 2008]

Definition (Named entities)

Textual units that are similar to proper names which meaning is built through the operation of reference.

Example

STRIP *BH2* *2%* *Nitrogen*

ONTORULE

Information extraction

Definition (Information extraction)

A set of techniques for automatically extracting structured information from unstructured textual documents

[Sundheim, 1992][Grishman and Sundheim, 1996]

Example

Who gives a conference, when and where?

Goals

- Extracting relations between entities
- Discovering concepts properties and relations [Hearst, 192] [Aussenac-Gilles and Jacques, 2008]
- Extracting business rules
- ...

Information extraction method

Information relies on extraction patterns that must be carefully designed to achieve good precision and recall.

Example

$$noun_1, noun_2, \ldots \textit{ and other } noun_3 \rightarrow noun_1 < noun_3$$

Hospitals, schools and other public buildings $\rightarrow$ *school* $<$ *public building*

Mr. X $\rightarrow$ X is an instance of MAN

Difficulties

- Extraction patterns are often corpus dependent
- Manual design is tedious and error prone
- Machine learning can help but requires large amount of annotated data to extract relational information
  [Califf M. E., 1998] [Brin, 1999]

Building business models from texts

Goal Combining state-of-the-art approaches

- To bootstrap and help business modeling
- To enable business analyst to author business models
- To deliver understandable business models
- To anchor business models in policies

Strategy Three major steps

- Designing an acquisition corpus
- Acquiring the domain conceptual model (domain ontology)
- Modeling business rules

Corpus design

Designing an acquisition corpus is a complex task [Atkins *et al.*, 1992]

- An acquisition corpus is application dependent
- Identifying relevant documents is challenging in many organizationions
- There exist a wide variety of source documents
  - Text types: educational material reflecting domain *vs.* working papers expressing factual knowledge
  - Corpus size: from 10 Kwords to 100 Mwords
  - Redundancy
  - Technicality
  - Target audience
- There may be missing pieces of knowledge
- Documents are often redundant
- Some documents may be confidential

How to achieve a good representativity of the domain to model?

In practice

- There is no stable methodology or acknowledged good practice to design acquisition corpus: all use cases are different.
- Knowledge engineers try to inventory the available documentation, from which they extract the most useful documents.

Example (ArcelorMittal use case)

- Extract from the product catalogue (10 pages, 3,500 words)
- Use case description (2,200 Kwords)
- Description of the order Assignment at galvanizationion Line (1,000 words)
- Email discussion with experts

Ontology design

The TERMINAE approach to ontology acquisition

[Aussenac-Gilles *et al.*, 2008][Szulman *et al.*, 2009b]

- An interactive approach
- A terminological approach

TERMINAE tool (http://www-lipn.univ-paris13.fr/ szulman/logi/index.html)

From texts to domain model

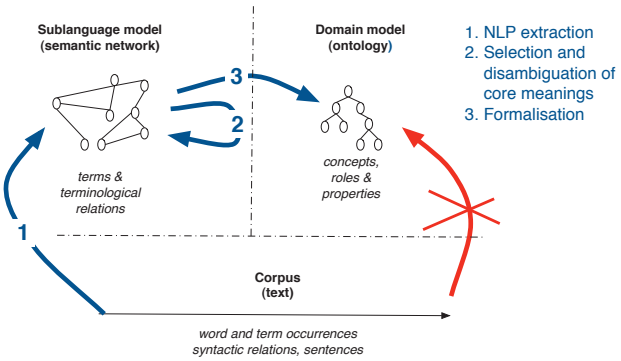
Texts present a view of the underlying domain model . . .

- Domain vocabulary: *hardening element, iron crystal lattice*
- Controlled meaning: *elements = chemical element*
- Situated communication (e.g. actors, locations, roles)
- Explicit domain knowledge: *The grain structure of steel influences its mechanical behavior*
- Factual knowledge: *Coil #13 is a coil, with length 670 meters, currently located in Aviles factor.*

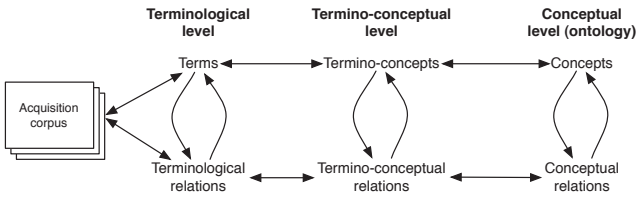
But a partial and distorted view

- Synonymy: *defects = non-conformities*
- Polysemy: *(lab test)* ***results*** $\neq$ ***result*** *(of the assignment process)*
- Ellipses
- Presuppositions

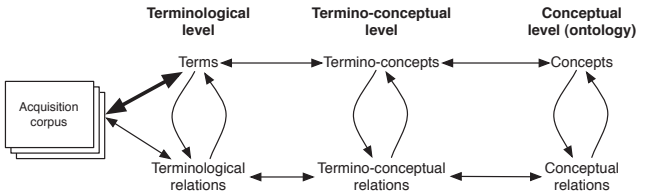
From a corpus to an ontology



Overall approach



1st step: Terminological analysis



Terminology extraction

Definition (Terminology)

The body of words or terms relating to a particular subject, field of activity or branch of knowledge.

Definition (term extractor)

Tools that take a domain specific acquisition corpus as input and output a list of specialized term candidates (*e.g.* YaTeA [Aubin and Hamon, 2006])

Underlying hypothesis

- The relevant terms of a corpus reflect the domain concepts
- Terminological analysis can bootstrap the ontology design

Terminology extraction and tagging

Example

The Galvanizationion Line processes coils, long strips of steel, to provide them a coating of zinc; this coating will give the product an improved surface aspect as well as protection from corrosion. During the process the mechanical properties, such as the yield strength, of the steel are also changed due to the thermal cycle it goes through.

Terminology extraction: methods

- Linguistic approach
  - Corpus tagging
  - Shallow syntactic analysis to identify phrase boundaries
  - Endogeneous disambiguation based on corpus redundancy
- Statistical approach
  - Repeated word sequences
  - Associated pairs of words
  - Syntactically filtered collocations
  - ...
- Mixed approach
  - Syntactically filtered collocations

[Daille *et al.*, 2004] [Pazienza *et al.*, 2005]

Linguistic approach

Example

Source text *The Galvanizationion Line processes coils [...] inducing residual carbon precipitation*

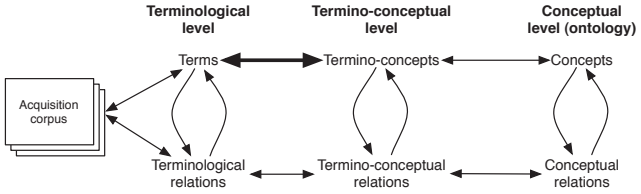
Tagged text The_{Det} $Galvanizationion_{Noun}$ $Line_{Noun}$ $processes_{Verb}$ $coils_{Noun}$ *[...]* $inducing_{Verb}$ $residual_{Adl}$ $carbon_{Noun}$ $precipitation_{Noun}$

Parsed text *The [Galvanizationion Line]*$_{NP}$ $processes_{Verb}$ $coils_{Noun}$ *[...]* $inducing_{Verb}$ *[residual carbon precipitation]*$_{NP}$

Disambiguation *[[residual carbon] precipitation]* or *[residual [carbon precipitation]]*?

Extracted terms *Galvanizationion Line, coils, residual carbon, residual carbon precipitation*

2nd step: Term normalizationion



Term normalizationion

Goal: abstracting from language peculiarities?

- Term filtering and selection (noise)
- Term variant clustering (synonymy)
- Term disambiguation (polysemy)

Output

- A semantic network of normalized terms and terminological relations
- TERMINAE can export the result in SKOS[1]

 [1] http://www.w3.org/2004/02/skos/

Term filtering and selection

Term extractors provide noisy results (candidate terms) that must be further filtered [Nazarenko and Zargayouna, 2009]

- Terminology creation is a social process: there is a consensus within a given community to use a term t with a specific meaning
- Termhood cannot be fully captured through linguistic and statistical properties
- Generic extractors cannot capture all the domain peculiarities

Validation interfaces allow for

- Character based filtering
- Lexical filtering
- Term ranking (ordered by frequency, tf.idf, *etc.*)