Domain specific facet based language resource

Biswanath Dutta

bisu@disi.unitn.it
DISI, University of Trento, Trento, Italy
Topics

- Introduction
- Language Resources (LR)
- Knowledge Organization Systems (KOS)
- DERA
- From DERA to Description Logics (DL)
- Demo: Modeling a Domain using DERA
- UniTn UK
- Conclusion and Future Work
Introduction

- Language resources
  - General purpose language resources
    - WordNet (http://wordnet.princeton.edu/)
    - EuroWordNet (http://www.illc.uva.nl/EuroWordNet/)
    - Roget’s thesaurus
  - Domain specific language resources
    - Dewey Decimal Classification (DDC)
    - Library of Congress Classification (LCC)
    - Universal Decimal Classification (UDC)
    - Bliss Bibliographic Classification (BC)
    - Colon Classification (CC)
    - AGROVOC
    - Art and Agriculture Thesaurus

- UniTn UK
Uses of the Language Resources

- Overall uses
  - Classification of contents of a library, the Internet or other databases (*for systematic organization and for easy management*)
  - Search of contents of a library, the Internet or other databases (*for retrieval*)

- Other uses
  - Natural Language Processing (NLP)
    - Automatic summarization
    - Named entity recognition
    - Word Sense Disambiguation
    - ...
Overall uses

- Classification
  - Document classification (e.g., book, journal article, news article, scientific papers, web page, etc.)
  - Items in a supermarket (e.g., detergents, food items, vegetables)
  - ...

- Search
  - Cataloguing
  - Indexing
**Other Uses**

Natural Language Processing

- **Automatic text summarization:**
  - Produce a readable **summary of a chunk of text**. Often used to provide summaries of text of a known type, for example, articles in the sport section of a newspaper.

- **Named entity recognition:**
  - Given a stream of text, determine which items in the text **map to proper names**, such as people or places, and what the type of each such name is (e.g. person, location, organization).

- **Word Sense Disambiguation:**
  - Many **words have more than one meaning**; we have to select the meaning which makes the most sense in context.

- ...

http://en.wikipedia.org/wiki/Natural_language_processing
Challenges

- Challenges we are facing in general
  - High recall
  - Low precision

- Natural language processing
  - Disambiguation problems
    - E.g., a word “bass”
      - Sense 1: A kind of saltwater fish
      - Sense 2: Tones of low frequency

  Natural language sentences:
  - I went fishing for some sea bass
  - The bass line of the song is too weak

- There is a lack of availability of large scale and high quality language resources
Solution

• A Large Scale, Domain Specific LR based on Facet based KO is a better Resource for addressing the challenges of Low Precision and High Recall all applications using LR like WordNet are facing.
Objective

- At the end of this presentation we will know the followings:
  - Fundamental differences between enumerative and facet based KOS systems
  - Modeling domain specific knowledge using DERA, a faceted knowledge organization framework
  - UniTn Universal Knowledge base (UK), a large scale high quality language resource
General Purpose Language Resources
WordNet

- is a large lexical database for the English language, developed at the Cognitive Science Laboratory at Princeton University
- it groups words of different part of speech (nouns, verbs, adjectives and adverbs) into sets of cognitive synonyms, called synsets, each expressing a distinct concept (i.e., each synset groups all the words with same meaning or sense)
- Synsets are interlinked by means of conceptual-semantic and lexical relations
- Typical semantic relations are hypernym (is-a) and part meronym (part-of)
- WordNet 2.1 consists of:
  - 117,597 synsets, 354,057 relations, 147,252 terms and 207,019 senses
MultiWordNet

- is a multilingual lexical database
- it includes languages such as, English, Italian, Spanish, Portuguese, Hebrew, Romanian, Latin
- Italian WordNet is strictly aligned with WordNet 1.6
- It consists of:
  - 102100+ synsets, 131300+ terms, 176600+ senses [English]
  - 38500+ synsets, 46200+ terms, 67400+ senses [Italian]
  - 57400+ synsets, 67200+ terms, 106500+ senses [Spanish]
  - 17200+ synsets, 16200+ terms, 21500+ senses [Portuguese]
  - 5900+ synsets, 0 terms, 0 senses [Hebrew]
  - 20100+ synsets, 20000+ terms, 34200+ senses [Romanian]
  - 8973+ synsets, 9100+ terms, 25900+ senses [Latin]
Domain Specific KOS as Language Resources
What is Knowledge Organization (KO)?

- It is about organization of knowledge, where organization refers to the arrangement and organizing structure to arrange and classify.

- KO = KO Processes (KOP) + KO Systems (KOS)

- KOP = *is about activities* such as document description, indexing and classification performed in libraries, databases, archives etc.

- KOS = *is about the tools that present the organized interpretation of knowledge structures* which include *classification* (and categorization) schemes that organize materials at a general level, *subject headings* that provide more detailed access, and *authority* files that control variant versions of key information such as geographic names and personal names. Also include highly structured vocabularies, such as *thesauri*, and *less traditional schemes*, such as semantic networks and ontologies (Hodge, 2000)

Library Classification Systems

- DDC: Enumerative classification system
- LCC
- CC
- BC: Faceted classification system
- UDC
Enumerative vs. Facet
Enumerative

- List all the classes of objects in a nested tree structure
- The classes are organized in successive, subordinate relationships
- In analogous to a simplified language where all possible statements have been pre-defined
  - E.g., French poetry, Italian drama, Chicken ribs

**Important:** it is impossible to list the universe of all possible subjects (a subject is a composition of one or more concepts)
- E.g., Literature, French poetry, Italian drama, Chicken ribs, Island
Enumerative Literature

English literature
  Poetry
  Drama
  Fiction

Italian literature
  Poetry
  Drama
  Fiction

Tamil literature
  Poetry
  Drama
  Fiction
Facet

[First Introduced by Ranganathan (1930s) in Library and Information Science]

• “A generic term used to denote any component – be it a basic subject or an isolate – of a compound subject, …” - Ranganathan

• It is a category that expresses some aspect of the knowledge being described

• A facet is a hierarchy of homogeneous terms, where each term in the hierarchy denotes a primitive atomic concept

• E.g., Organ facet, geographical facet, language facet, property facet, author facet, religion facet, commodity facet, etc.
Facet Example

**Language**

by Indo-European
- Teutonic
  - Gothic
- English
  - American English
  - German
- Latin
  - Italian
- French
- Greek

by Dravidian
- Tamil
- Tulu

by Geographic location
- Asian language
  - (collective treatment)
  - Japanese language
  - Indian language
  - African language
In analogy to natural language, it is like a vocabulary of nouns, verbs, adjectives, etc., where the words and the types of words have been defined, but the speakers combine them as needed to create statements.
Characteristics of Facet based Systems

- Flexibility
- Expandability
- Reusability
- Compactness
- Robustness
- Consistency
Facet based KO Frameworks
PMEST

[Ranganathan, 1933]

- Consists of:
  - Personality [P]
  - Matter [M] \( (\text{matter material, matter method, matter property}) \)
  - Energy [E] \( (\text{internal and external actions, processes}) \)
  - Space [S]
  - Time [T]


- **Drawbacks:**
  - Sometime it is difficult to interpret, particularly the “P”
  - No explicit relations (is-a/part-of) between the classes
  - More syntactic approach than semantic
  - Does not allow for direct translation of the elements of its facets P, M, E, S, T into DL axioms
  - No clear separation between entity and entity classes
  - ...
DEPA

[G. Bhattacharyya, 1979]
- It is a simplified version of Ranganathan’s PMEST
- Consists of:
  - Domain [D]
  - Entity [E]
  - Property [P]
  - Action [A]
  - Plus one additional facet: Modifier [m]
    - Modifiers: space, time, form, language, environment, etc.

- **Drawbacks:**
  - No explicit relations (is-a/part-of) between the classes
  - More syntactic approach than semantic
  - Does not allow for direct translation of the elements of its facets E, P, A into DL axioms
  - No clear separation between entity and entity classes
  - ...
DERA

[F. Giunchiglia and B. Dutta, 2011]

- Consists of:
  - Domain [D]
  - Entity [E]
  - Relation [R]
  - Attribute [A]

- It is a further refined and simplified form of Bhattacharyya’s DEPA

- Has direct mapping to DL

- Emphasis is on the named entities
DERA

- A facet based KO Framework
- A framework independent of any particular domain
- A framework for developing domain specific knowledge resources
- Has mapping to Description Logics (DL)
- Is logically sound
- Is designed in UniTn KnowDive group
Domain

- An area of knowledge that we are interested in or we are communicating about it
- An organized field of knowledge that deals with specific kinds of subjects
- E.g., Space, Time, Music, Movie, Sport, Mathematics, etc.
DERA (2)

- **D = <E, R, A>**
  - **E = Entity** - an elementary component consisting of classes and their instances, having either perceptual correlates or only conceptual existence within a domain in context.
    - E.g., in the Space domain, natural elevations, such as, mountain, hill, seamount etc. are entity classes, while the Himalaya, Monte Bondone, Loihi seamount etc. are entities
  - **R = Relation** - an elementary component consisting of classes representing the *relation* between entities.
    - E.g., in the Space domain, north, south, near, adjacent, in front, etc. are spatial relations between entities
  - **A = Attribute** - an elementary component consisting of classes denoting the *qualitative/quantitative* or *descriptive* properties of entities.
    - E.g., in the Space domain, altitude (of a hill), length (of a river), surface area (of a lake), etc. are qualitative/quantitative properties, while the kinds of rocks (of a mountain), architectural style (of a monument) are descriptive attributes
An elementary component that consists of classes (categories) and their instances, having either perceptual correlates or only conceptual existence in a domain in context.

- $E = \langle \{e\}, \{\mathcal{E}\} \rangle$
- $e = \text{Entity class} \cdot \text{consists of the core classes within a domain}$
- $\mathcal{E} = \text{Entity} \cdot \text{consists of the real world (named) entities which are instances of the entity classes “e”}$
Relation

- An elementary component that consists of relations inside a domain

- \[ R = \langle \{ r \} \rangle \]

- \( r = \text{Relation} \) – consists of the classes representing the relations between entities

- Relations play an important role for effective knowledge discovery, for example,
  - Retrieve all the secondary schools within 500 meters of the Dante railway station in Trento
  - Find all the highways of the Trentino province adjacent to marine areas
An elementary component consists of classes belonging to or characteristic of entities.

\[ A = \langle \mathcal{A}, \mathcal{E} \rangle \]

\( \mathcal{A} \) = Datatype attribute – consists of classes which qualify or quantify the entities

\( \mathcal{E} \) = Descriptive attribute – consists of classes describing entities
From DERA to Description Logics (DL)
Mapping to DL

- From DERA to DL
  - Entity class (e) -> Concepts
  - Relation (R) -> Roles
  - Datatype attribute (A) -> Roles
  - Descriptive attribute (e) -> Roles
  - Entity (E) -> Individuals
Mapping

- An Interpretation of a DERA domain consists of:
  - an Interpretation Function \( I \) and a not empty set \( D' \) (the domain of Interpretation)

- \( D' \) consists of set of:
  - Concepts \( e' \)
  - Entities \( E' \)
  - Relations \( R' \)
  - Datatype attributes \( A' \)
  - Descriptive attributes \( \varphi' \)

\[
e' \subseteq D', \quad E' \subseteq D', \quad R' \subseteq D' \times D', \quad A' \subseteq D' \times D', \quad \varphi' \subseteq D' \times D'
\]
# TBox and ABox

<table>
<thead>
<tr>
<th>TBox</th>
<th>ABox</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flowing body of water (\subseteq) Body of water</td>
<td>instance_of(Lake Garda, Lake)</td>
</tr>
<tr>
<td>Natural flowing body of water (\subseteq) Flowing body of water</td>
<td>near(Lake garda, Lake Caldonazzo)</td>
</tr>
<tr>
<td>Stream (\subseteq) Natural flowing body of water</td>
<td>near(Lake Garda, Venice)</td>
</tr>
<tr>
<td>Internal spatial relation (\subseteq) Spatial relation</td>
<td>central(Duomo, Trento)</td>
</tr>
<tr>
<td>Central (\subseteq) Internal spatial relation</td>
<td>architectural Style(Colosseum, Roman architecture)</td>
</tr>
<tr>
<td>Midplane (\subseteq) Central</td>
<td>primaryInflow(Gordan, Great salt lake)</td>
</tr>
<tr>
<td>Volume (\subseteq) Dimension</td>
<td></td>
</tr>
<tr>
<td>Roman architecture (\subseteq) Architectural style</td>
<td></td>
</tr>
</tbody>
</table>

**Inferred facts**, e.g.,

1. “Midplane” is an “Internal spatial relation”
2. “Lake Caldonazzo” is near “Venice”
Demo: Modeling a Domain Specific Knowledge
Analytico-Synthetic Approach

[Ranganathan (1933)]

- A well established technique for building classificatory structures from atomic concepts which are analysed into facets and sequenced according to the syntax of a system.

- In this approach a domain under examination is decomposed into its basic terms and organize those terms into facets to display the relationships between them.

- A faceted representation scheme codifies the basic building blocks that can be used - at indexing, classification and searching time, and also to construct complex labels (subjects).

  “Stormy weather [topic] during the rainy season [time] in the Java Island [space]”
Methodology

- Step 1: Identification of the atomic concepts
- Step 2: Analysis (per genus et differentiam)
- Step 3: Synthesis
- Step 4: Standardization
- Step 5: Ordering

Following the above steps leads to the creation of a set of facets. They constitute a faceted representation scheme for a domain.
Principles

1. **Relevance** (e.g., *breed* is more realistic to classify the universe of cows instead of by *grade*)

2. **Ascertainability** (e.g., flowing body of water)

3. **Permanence** (e.g., Spring- a natural *flow* of ground water)

4. **Exhaustiveness** (e.g., to classify the universe of *people*, we need both *male* and *female*)

5. **Exclusiveness** (e.g., *age* and *date of birth*, both produce the same divisions)

6. **Context** (e.g., *bank*, a bank of a river, OR, a building of a financial institution)
   - Important: helps in reducing the homographs

7. **Currency** (e.g., metro station vs. subway station)

8. **Reticence** (e.g., minority author, black man)

9. **Ordering**
   - Important: ordering carries semantics as it provides implicit relations between the coordinate terms
Domain: Painting

- **Step 1:**

- **Identification of the atomic concepts (from various sources):**

  - Human figure, Bust, Earth, Mountain, Water, Lake, Sky, Plant, Animal, Painter, Location, Year, Red, Orange, Blue, height, width, Oil paint, Water-based paint, Distemper, History, ...
Domain: Painting

- **Step 2:** Analysis (and categories the concepts based upon their dissimilar and similar characteristics)
  - {Human figure, Bust}
  - {Earth, Mountain, Water, Lake, Sky, Plant, Animal}
  - Painter
  - Location
  - Year
  - {Red, Orange, Blue}
  - {Height, Width}
  - {Oil paint, Water-based paint, Distemper}
  - History
Domain: Painting

- **Step 3**: Synthesis (build the facets and name/label the categories)

- **Figure** \{Human figure(Bust)}

- **Nature** \{(Earth, Mountain, Water, Lake, Sky), Plant, Animal\}

- Painter

- Location

- Year

- **Color** \{Red, Orange, Blue\}

- **Dimension** \{Height, Width\}

- **Painting type** \{Oil paint, Water-based paint, Distemper\}

- History
Domain: Painting

- **Step 4**: standardize the terms
- **Step 5**: model the domain according to the DERA structure

**Entity class (e):**
- Figure
  - Human figure
  - Bust
- Nature
  - Earth
  - Mountain
  - Water
  - Lake
- Sky
- Plant
- Animal

**Relation (R):**
- Painter
- Location
- Year

**Datatype attribute ($\mathcal{A}$):**
- Color
  - Red
  - Orange
  - Blue
- Dimension
  - Height
  - Width

**Descriptive attribute ($\mathcal{E}$):**
- Painting type
  - Oil paint
  - Water-based paint
  - Distemper
- History

Background Knowledge (BK) for “Painting” domain
Advantages of BK

- Projection of the domain specific knowledge
- Categorization is **not** merely limited to the entity types, but also includes the relations between entities (i.e. interactions with other entities), behavior of the entities, process in it or on it, and so on
- All the related knowledge necessary to express a domain is in one place
- In a domain, a word can have only one meaning (concept)
UniTn UK
A knowledge base, developed by KnowDive group, University of Trento

Allows encoding knowledge in multiple languages

Started with WordNet 2.1 + MultiWordNet

Also, imported knowledge from:
- GeoNames (http://www.geonames.org/), PAT (Autonomous Province of Trento), YAGO (http://www.mpi-inf.mpg.de/yago-naga/yago/), etc.

300+ domains

At present it consists of:
- 150,000+ terms
- 10,000,000 entities
- 93,000,000+ axioms
UniTn UK Structure

INTRODUCTION::GPLD::KOS::DERA::MAPPING TO DL::DEMO::UNITN UK::CONCLUSION

UK

Linguistic resources (e.g. WordNet) + KOS (e.g. thesauri, classification schemes)

Integration

Representation
Reasoning
Maintenance
Evolution

BACKGROUND KNOWLEDGE
(personalization)

LINGUISTIC KNOWLEDGE
(words and senses in different languages)

ENTITIES

CONCEPTS AND RELATIONS

DOMAIN KNOWLEDGE
UK and DERA

In this block domain independent knowledge is structured in facets*

Domains are modeled in this block according to the DERA framework

In building domain specific knowledge, we refer to the facets/ sub-facets defined in the “CONCEPTS AND RELATIONS” block

*Note that, at the beginning we populated this block by uploading the concepts and relations from WordNet 2.1 and latter on we applied the analytico-synthetic approach to build the facets
Conclusion and Future Work

- General purpose and domain specific language resources
- State-of-the-art problems
- Enumerative and Faceted approach in KOS
- KO frameworks
- How to model a domain using DERA
- UniTn UK

**Future work:**
- Our aim is to add more knowledge into UK in terms of terms, axioms and entities
- Our aim is to implement more number of domains
- Our aim is to develop UK as multilingual resources. At present UK knowledge base is in English language and a very limited set of knowledge is in Italian language
Further Readings

- V. Broughton, 2006. The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings, 58(1/2) pp. 49-72.