Generating Semantic Annotations

Anna Fensel
31.10.2016

Where are we?

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>Semantic Web Architecture</td>
</tr>
<tr>
<td>3</td>
<td>Resource Description Framework (RDF)</td>
</tr>
<tr>
<td>4</td>
<td>Web of data</td>
</tr>
<tr>
<td>5</td>
<td>Generating Semantic Annotations</td>
</tr>
<tr>
<td>6</td>
<td>Storage and Querying</td>
</tr>
<tr>
<td>7</td>
<td>Web Ontology Language (OWL)</td>
</tr>
<tr>
<td>8</td>
<td>Rule Interchange Format (RIF)</td>
</tr>
<tr>
<td>9</td>
<td>Reasoning on the Web</td>
</tr>
<tr>
<td>10</td>
<td>Ontologies</td>
</tr>
<tr>
<td>11</td>
<td>Social Semantic Web</td>
</tr>
<tr>
<td>12</td>
<td>Semantic Web Services</td>
</tr>
<tr>
<td>13</td>
<td>Tools</td>
</tr>
<tr>
<td>14</td>
<td>Applications</td>
</tr>
</tbody>
</table>
Agenda

• Motivation
• Technical solution, illustrations, and extensions
  – Semantic annotation of text
  – Semantic annotation of multimedia
  – Annotation with schema.org
• Large example
• Summary
• References

MOTIVATION
Semantic Annotation

• Creating semantic labels within documents for the Semantic Web.

• Used to support:
  – Advanced searching (e.g. concept)
  – Information Visualization (using ontology)
  – Reasoning about Web resources

• Converting syntactic structures into knowledge structures

Semantic Annotation Process
Manual semantic annotation

- Manual annotation is the transformation of existing syntactic resources into interlinked knowledge structures that represent relevant underlying information.

- Manual annotation is an expensive process, and often does not consider that multiple perspectives of a data source, requiring multiple ontologies, can be beneficial to support the needs of different users.

- Manual annotation is more easily accomplished today, using authoring tools such as Semantic Word:

  [BAGRAM], Afghanistan (CNN) -- British commandos have unearthed a large weapons cache inside a "significant" al Qaeda cave complex in Afghanistan,

  UK military officials say

  The cave system is still being searched but anti-aircraft, anti-tank and other heavy munitions have already been found.
Semi-automatic semantic annotation

- Semi-automatic annotation systems rely on human intervention at some point in the annotation process.

- The platforms vary in their architecture, information extraction tools and methods, initial ontology, amount of manual work required to perform annotation, performance and other features, such as storage management.

- Example: GATE (see in section 2.1 and 3).

Automatic semantic annotation

- Automatic semantic annotation is based on the automatic annotating algorithms: e.g., PANKOW (Pattern-based Annotation through Knowledge On the Web), C-PANKOW (Context-driven and Pattern-based Annotation through Knowledge on the Web) for texts; statistical algorithms for image and video annotations.

- However, annotations based on automatic algorithms mostly need to be proved and corrected after implementation of these algorithms.

- EXAMPLE of tools: OntoMat can provide fully automated annotation and interactive semi-automatic annotation of texts.
- M-OntoMat is an automatic multimedia annotation tool (see 2.2 Multimedia Annotation).
- ALIPR is a real-time automatic image tagging engine.
Automatic semantic annotation: OntoMat

• OntoMat-Annotizer was created by S. Handshuh, M. Braun, K. Kuehn, L. Meyer within OntoAgent project

• OntoMat supports two modes of interaction with PANKOW-algorithm: (1) fully automatic annotation, and (2) interactive semi-automatic annotation.

• In the fully automatic mode, all categorizations with strength above a user-defined are used to annotate the Web content.

• In the interactive mode, the system proposes the top five concepts to the user for each instance candidate. Then, the user can disambiguate and resolve ambiguities (see the illustration below).
Automatic semantic annotation: ALIPR

- ALIPR stands for "Automatic Linguistic Indexing of Pictures—Real Time"
- It is an Automatic Photo Tagging and Visual Image Search
- ALIPR was developed in 2005 at Pennsylvania State University by Professors Jia Li and James Z. Wang and was published and made public in October 2006.
- ALIPR version 1.0 is designed only for color photographic images.
- After writing in the URL or after image upload, the tool automatically offers the tags for the image annotation (see illustration with a flower in the next slide)
Automatic semantic annotation: ALIPR

- ALIPR annotates images based on content.

- First, it learnt to recognize the meaning of the tags before suggesting the correct labels. As part of the learning process, the researchers fed ALIPR hundreds of images of the same topic, for example “flower”. ALIPR analyzed the pixels and extracted information related to color and texture. It then stored a mathematical model for “flower” based on the cumulative data.

- Later, when a user uploads a new picture of a flower, ALIPR compares the pixel information from the pre-computed models in its knowledge base and suggests a list of 15 possible tags.

Semantic Annotation Concerns

- Scale, Volume
  - Existing & new documents on the Web
  - Manual annotation
    - Expensive – economic, time
    - Subject to personal motivation
    - Schema Complexity

- Storage
  - support for multiple ontologies
  - within or external to source document?
  - Knowledge base refinement

- Access - How are annotations accessed?
  - API, custom UI, plug-ins
Technical solution

2.1 Annotation of text
   • Semi-automatic text annotation
   • GATE
   • KIM

2.2 Multimedia annotation
   • Levels of multimedia annotation
   • Tools for multimedia annotation
   • Multimedia ontologies
   • „Games with a purpose“

2.3 Annotation with schema.org
   • Vocabulary for annotation
   • Tools and examples
Annotation of text

- Many systems apply rules or wrappers that were manually created that try to recognize patterns for the annotations.
- Some systems learn how to annotate with the help of the user.
- Supervised systems learn how to annotate from a training set that was manually created beforehand.
- Semi-automatic approaches often apply information extraction technology, which analyzes natural language for pulling out information the user is interested in.
A Walk-Through Example: GATE

GATE is a leading NLP and IE platform developed in the University of Sheffield, consists of different modules:

- Tokeniser
- Gazetteer
- Sentence Splitter
- Part-of-Speech Tagger (POS-Tagger)
- Named Entity Recogniser (NE-Recognizer)
- OrthoMatcher (Orthographic Matcher)
- Coreference Resolution

Tokeniser

The tokeniser splits the text into very simple tokens such as numbers, punctuation and words of different types:

<table>
<thead>
<tr>
<th>Type</th>
<th>Set</th>
<th>Start</th>
<th>End</th>
<th>Id</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>0</td>
<td>4</td>
<td>749</td>
<td>6</td>
<td>(inewed, length=4, orthnumeral, orth=Gold)</td>
</tr>
<tr>
<td>Token</td>
<td>5</td>
<td>8</td>
<td>751</td>
<td>9</td>
<td>(inewed, length=4, orthlowercase, orth=loos)</td>
</tr>
<tr>
<td>Token</td>
<td>10</td>
<td>11</td>
<td>763</td>
<td>12</td>
<td>(inew-symbol, length=1, string-3, symbkind-currency)</td>
</tr>
<tr>
<td>Token</td>
<td>11</td>
<td>12</td>
<td>764</td>
<td>13</td>
<td>(inewnumber, length=1, string=1)</td>
</tr>
<tr>
<td>Token</td>
<td>12</td>
<td>13</td>
<td>765</td>
<td></td>
<td>(inewpunctuation, length=1, string=,)</td>
</tr>
<tr>
<td>Token</td>
<td>13</td>
<td>14</td>
<td>766</td>
<td>19</td>
<td>(inewnumber, length=2, string=23)</td>
</tr>
<tr>
<td>Token</td>
<td>16</td>
<td>17</td>
<td>767</td>
<td>17</td>
<td>(inew-punctuation, length=1, string=,)</td>
</tr>
<tr>
<td>Token</td>
<td>18</td>
<td>22</td>
<td>769</td>
<td>22</td>
<td>(inewnumber, length=4, orthlowercase, string=hits)</td>
</tr>
<tr>
<td>Token</td>
<td>23</td>
<td>28</td>
<td>761</td>
<td>38</td>
<td>(inewnumber, length=5, orthlowercase, string=hits)</td>
</tr>
</tbody>
</table>
Semantic Gazetteer Lookup

The gazetteer lists used are plain text files, with one entry per line. Each list represents a set of names, such as names of cities, organizations, days of the week, etc.

Sentence Splitter

The sentence splitter is a cascade of finite-state transducers which segments the text into sentences. This module is required for the tagger. The splitter uses a gazetteer list of abbreviations to help distinguish sentence-marking full stops from other kinds.
Part-of-Speech Tagger (POS-Tagger)

- POS-Tagger produces a part-of-speech tag as an annotation on each word or symbol.
- Neither the splitter nor the tagger are a mandatory part of the IE system, but the extra linguistic information they produce increases the power and accuracy of the IE tools.

Ontology-aware NER (Named Entity Recogniser)

The named entity recogniser consists of pattern-action rules, executed by the finite-state transduction mechanism. It recognizes entities like person names, organizations, locations, money amounts, dates, percentages, and some types of addresses.

In order to illustrate the process, consider a snippet of text. The recogniser first identifies entities such as proper names, dates, and numbers, which it then uses to populate a knowledge base. This base can be used for subsequent analysis and processing.

The recogniser also handles special cases, like abbreviations and contractions, ensuring that even complex texts are accurately parsed.

Finally, the recogniser outputs a structured representation of the text, with each entity clearly marked and categorized. This output can then be used for further analysis, such as sentiment analysis or named entity resolution.
OrthoMatcher = Orthographic Coreference

- The OrthoMatcher module adds identity relations between named entities found by the semantic tagger, in order to perform coreference.
- The matching rules are only invoked if the names being compared are both of the same type, i.e. both already tagged as (say) organizations, or if one of them is classified as ‘unknown’. This prevents a previously classified name from being re-categorized.

Pronominal Coreference Resolution

- quoted text submodule
- pleonastic it submodule
- pronominal resolution submodule
Quoted Text Submodule

The quoted speech submodule identifies quoted fragments in the text being analyzed. The identified fragments are used by the pronominal coreference submodule for the proper resolution of pronouns such as I, me, my, etc. which appear in quoted speech fragments.

"There's like elixirs in a bottle," Boehner said at the POS. "It's always difficult to get the first one out but then the rest come easily."

Besides the expected stock offerings, Schwarzman said Blackstone in recent months has agreed to sell four additional companies in its portfolio, and is tripling the sale of a fifth.

"Vehicular" alone a company would pump $2.8 billion into Blackstone's investment funds, he said.

Blackstone is also scoring new investment opportunities after the dry spell wrought by the credit crisis. Schwarzman and lenders are again starting to provide loans to new deals — if not at the levels they did before the market soared.

"I think there are signs of life in the bank market," he said. "We can certainly do transactions in the $3 to $4 billion range."

<table>
<thead>
<tr>
<th>Type</th>
<th>Start</th>
<th>End</th>
<th>Id</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>QuotedText</td>
<td>1518</td>
<td>1541</td>
<td>1340</td>
<td>(needy-bole-01)</td>
</tr>
<tr>
<td>QuotedText</td>
<td>2611</td>
<td>2639</td>
<td>1340</td>
<td>(needy-bole-01)</td>
</tr>
</tbody>
</table>

Pleonastic "It" Submodule

The pleonastic "it" submodule matches pleonastic occurrences of "It". Similar to the quoted speech submodule, it is a transducer operating with a grammar containing patterns that match the most commonly observed pleonastic "it" constructs.

"Looks as if the bottom has been reached," he said, adding that "the world's financial system is in a radically different place" from its precarious state a year ago.

<table>
<thead>
<tr>
<th>Type</th>
<th>Start</th>
<th>End</th>
<th>Id</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleonastic</td>
<td>1767</td>
<td>1780</td>
<td>1342</td>
<td>(needy-bole-01)</td>
</tr>
</tbody>
</table>
Pronominal Coreference Resolution

The main functionality of the coreference resolution module is in the pronominal resolution submodule. This module finds the antecedents for pronouns and creates the coreference chains from the individual anaphor/antecedent pairs and the coreference information supplied by the OrthoMatcher.

KIM platform

- KIM = Knowledge and Information Management
- developed by semantic technology lab „Ontotext“
- based on GATE
KIM platform

- KIM performs IE based on an ontology and a massive knowledge base.

KIM KB

- KIM KB consists of above 80,000 entities (50,000 locations, 8,400 organization instances, etc.)

- Each location has geographic coordinates and several aliases (usually including English, French, Spanish, and sometimes the local transcription of the location name) as well as co-positioning relations (e.g. `subRegionOf`.)

- The organizations have `locatedIn` relations to the corresponding Country instances. The additionally imported information about the companies consists of short description, URL, reference to an industry sector, reported sales, net income, and number of employees.
The KIM platform provides a novel infrastructure and services for:

- automatic semantic annotation,
- indexing,
- retrieval of unstructured and semi-structured content.

The most direct applications of KIM are:

- Generation of meta-data for the Semantic Web, which allows hyper-linking and advanced visualization and navigation;
- Knowledge Management, enhancing the efficiency of the existing indexing, retrieval, classification and filtering applications.
KIM platform

- The automatic semantic annotation is seen as a named-entity recognition (NER) and annotation process.
- The traditional flat NE type sets consist of several general types (such as Organization, Person, Date, Location, Percent, Money). In KIM the NE type is specified by reference to an ontology.
- The semantic descriptions of entities and relations between them are kept in a knowledge base (KB) encoded in the KIM ontology and residing in the same semantic repository. Thus KIM provides for each entity reference in the text (i) a link (URI) to the most specific class in the ontology and (ii) a link to the specific instance in the KB. Each extracted NE is linked to its specific type information (thus Arabian Sea would be identified as Sea, instead of the traditional – Location).

KIM platform

KIM plug-in for the Internet Explorer browser
Multimedia Annotation

• Different levels of annotations
  – Metadata
    • Often technical metadata
    • EXIF, Dublin Core, access rights
  – Content level
    • Semantic annotations
    • Keywords, domain ontologies, free-text
  – Multimedia level
    • low-level annotations
    • Visual descriptors, such as dominant color
Metadata

• refers to information about technical details
• creation details
  – creator, creationDate, …
  – Dublin Core
• camera details
  – settings
  – resolution
  – format
  – EXIF
• access rights
  – administrated by the OS
  – owner, access rights, …

Content Level

• Describes what is depicted and directly perceivable by a human
• usually provided manually
  – keywords/tags
  – classification of content
• seldom generated automatically
  – scene classification
  – object detection
• different types of annotations
  – global vs. local
  – different semantic levels
Global vs. Local Annotations

- Global annotations most widely used
  - flickr: tagging is only global
  - organization within categories
  - free-text annotations
  - provide information about the content as a whole
  - no detailed information
- Local annotations are less supported
  - e.g. flickr, PhotoStuff allow to provide annotations of regions
  - especially important for semantic image understanding
    - allow to extract relations
    - provide a more complete view of the scene
  - provide information about different regions
  - and about the depicted relations and arrangements of objects

Semantic Levels

- Free-Text annotations cover large aspects, but less appropriate for sharing, organization and retrieval
  - Free-Text Annotations probably most natural for the human, but provide least formal semantics
- Tagging provides light-weight semantics
  - Only useful if a fixed vocabulary is used
  - Allows some simple inference of related concepts by tag analysis (clustering)
  - No formal semantics, but provides benefits due to fixed vocabulary
  - Requires more effort from the user
- Ontologies
  - Provide syntax and semantic to define complex domain vocabularies
  - Allow for the inference of additional knowledge
  - Leverage interoperability
  - Powerful way of semantic annotation, but hardly comprehensible by "normal users"
Tools

- Web-based Tools
  - flickr
  - riya
- Stand-Alone Tools
  - PhotoStuff
  - AktiveMedia
- Annotation for Feature Extraction
  - M-OntoMat-Annotizer

flickr

- Web2.0 application
- tagging photos globally
- add comments to image regions marked by bounding box
- large user community and tagging allows for easy sharing of images
- partly fixed vocabularies evolved
  - e.g. Geo-Tagging
riya

- Similar to flickr in functionality
- Adds automatic annotation features
  - Face Recognition
    - Mark faces in photos
    - associate name
    - train system
    - automatic recognition of the person in the future

PhotoStuff

- Java application for the annotation of images and image regions with domain ontologies
- Used during ESWC2006 for annotating images and sharing metadata
- Developed within Mindswap
AktiveMedia

- Text and image annotation tool
- Region-based annotation
- Uses ontologies
  - suggests concepts during annotation
  - providing a simpler interface for the user
- Provides semi-automatic annotation of content, using
  - Context
  - Simple image understanding techniques
  - flickr tagging data

M-OntoMat-Annotizer

- Extracts knowledge from image regions for automatic annotation of images
- Extracting features:
  - User can mark image regions manually or using an automatic segmentation tool
  - MPEG-7 descriptors are extracted
  - Stored within domain ontologies as prototypical, visual knowledge
- Developed within aceMedia
- Currently Version 2 is incorporating
  - true image annotation
  - central storage
  - extended knowledge extraction
  - extensible architecture using a high-level multimedia ontology
Multimedia Ontologies

- Semantic annotation of images requires multimedia ontologies
  - several vocabularies exist (Dublin Core, FOAF)
  - they don’t provide appropriate models to describe multimedia content sufficiently for sophisticated applications
- MPEG-7 provides an extensive standard, but especially semantic annotations are insufficiently supported
- Several mappings of MPEG-7 into RDF or OWL exist
  - now: VDO and MSO developed within aceMedia
  - later: Engineering a multimedia upper ontology

aceMedia Ontology Infrastructure

- aceMedia Multimedia Ontology Infrastructure
  - DOLCE as core ontology
  - Multimedia Ontologies
    - Visual Descriptors Ontology (VDO)
    - Multimedia Structures Ontology (MSO)
    - Annotation and Spatio-Temporal Ontology augmenting VDO and MSO
  - Domain Ontologies
    - capture domain specific knowledge
Visual Descriptors Ontology

- Representation of MPEG-7 Visual Descriptors in RDF
  - Visual Descriptors represent low-level features of multimedia content
  - e.g. dominant color, shape or texture
- Mapping to RDF allows for
  - linking of domain ontology concepts with visual features
  - better integration with semantic annotations
  - a common underlying model for visual and semantic features

Visual Knowledge

- Used for automatic annotation of images
- Idea:
  - Describe the visual appearance of domain concepts by providing examples
  - User annotates instances of concepts and extracts features
  - features are represented with the VDO
  - the examples are then stored in the domain ontology as prototype instances of the domain concepts
- Thus the names: prototype and prototypical knowledge
Extraction of Prototype

Transformation to VDO
Using Prototypes for Automatic Labelling

Knowledge Assisted Analysis

Multimedia Structure Ontology

- RDF representation of the MPEG-7 Multimedia Description Schemes
- Contains only classes and relations relevant for representing a decomposition of images or videos
- Contains Classes for different types of segments
  - temporal and spatial segments
- Contains relations to describe different decompositions
- Augmented by annotation ontology and spatio-temporal ontology, allowing to describe
  - regions of an image or video
  - the spatial and temporal arrangement of the regions
  - what is depicted in a region
Games with a purpose

Are proposed to masquerade the core tasks of weaving the Semantic Web behind online, multi-player game scenarios, in order to create proper incentives for human users to get involved.

Pioneer work: Luis von Ahn „Games with a purpose“

Games for semantic annotations:
ESP Game: Annotating Images

OntoTube: Annotating YouTube
OntoPronto: Annotating Wikipedia

Boston Grammar School

Boston Grammar School is a selective school for boys aged 11 to 18, recently admitting girls aged 16 to 18, in Boston, Lincolnshire. The school was founded by charter of Philip and Mary in 1520. The oldest part of the school was built in 1607 and formerly referred to as the "big school", is now used as the school library. Boston Grammar School was the model for Boston Latin School, the first.

Your choice

Hint: "Dog" is a class (as several entities of this class exist) "Lassie" is an instance

Starting singleplayer mode ...

Annotation with Schema.org
Schema.org Data Model

• Derived from RDFS
  • Some extensions now however go into higher expressivity e.g. of OWL

• Based on:
  • Set of Types (classes)
    • Organized in a hierarchy
    • Each type (class) might be a sub-class of several types (classes)

• Properties
  • Each property can have 1 or more items as domains
  • Each property can have 1 or more items as range

Data Model

• Canonical representation in RDFa
  • http://schema.org/docs/schema_org_rdfa.html

• Schema.org can be extended

• Schema.org properties can be used in other contexts

• The type hierarchy presented in Schema.org is not intended to be a 'global ontology' of the world.
Schema.org vocabularies

- Most popular vocabularies relates to...
  - CreativeWork
    - Book, Movie, Recipe, TVSeries, Review...
  - Embedded non-text objects: AudioObject, ImageObject,
  - Event
    - Food Event, Dance Event, Festival, SportsEvent...
  - Organization
  - Person
  - Place, Local Business, Hotel, Restaurant ...
  - Product, Offer

- All types of vocabularies can be found in: [http://schema.org/docs/full.html](http://schema.org/docs/full.html)

Schema.org vocabularies

- Support the following DataTypes
  - Boolean
    - False
    - True
  - Date
  - Date Time
  - Number
    - Float
    - Integer
  - Text
    - URL
  - Time
• For each item, Schema.org describes:
  • A list of own properties, range (datatype or item) and description
  • A list of inherited properties
  • A list of properties for which instances of the selected item may appear as values
  • A list of subclasses (more specific types)
  • Example of usage
How to mark-up with schema.org?

- Schema.org can be used to enrich the web sites with the following formats:
  - Microdata (most popular)
    - Tags introduced within HTML 5
    - Based on Item descriptions
    - Itemscope, Itemtype, Itemprop
  - RDFa
  - JSON-LD

Example I

Vocabulary – schema.org

- Example*:
  - Imagine you have a page about the movie Avatar—a page with a link to a movie trailer, information about the director, and so on. Your HTML code might look something like this:

```html
<div>
  <h1>Avatar</h1>
  <span>Director: James Cameron (born August 16, 1954)</span>
  <span>Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

* http://schema.org/docs/gs.html
Example I

- **Thing > Creative Work > Movie**
  - Particular properties

**Thing > CreativeWork > Movie**

A movie:

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actor</td>
<td>Person</td>
<td>An actor, e.g. in tv, radio, movie, video games etc. Actors can be associated with individual items or with a series, episode, clip. Supersedes actors.</td>
</tr>
<tr>
<td>Director</td>
<td>Person</td>
<td>A director of e.g. tv, radio, movie, video games etc. content. Directors can be associated with individual items or with a series, episode, clip. Supersedes directors.</td>
</tr>
<tr>
<td>Duration</td>
<td>Duration</td>
<td>The duration of the item (movie, audio recording, event, etc.) in ISO 8601 date format. The composer of the soundtrack.</td>
</tr>
<tr>
<td>Soundtrack</td>
<td>MusicGroup or Person</td>
<td>The production company or studio responsible for the item e.g. series, video game, episode etc.</td>
</tr>
<tr>
<td>Trailer</td>
<td>VideoObject</td>
<td>The trailer of a movie or tv/series, season, episode, etc.</td>
</tr>
</tbody>
</table>

Inherited properties (from Creative Work and Thing)

<table>
<thead>
<tr>
<th>Property from CreativeWork</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>The title of the content. Indicates that the resource is compatible with the referenced vocabulary or schema. A specific instance of this is a person.</td>
</tr>
<tr>
<td>Creator</td>
<td>The creator of the content. The person who actually made the content.</td>
</tr>
<tr>
<td>Contributor</td>
<td>The contributor of the content. The person who contributed to the content.</td>
</tr>
<tr>
<td>Publisher</td>
<td>The publisher of the content. The organization responsible for the publication of the content.</td>
</tr>
<tr>
<td>Source</td>
<td>The source of the content. The original file or the source of the content.</td>
</tr>
<tr>
<td>Language</td>
<td>The language of the content. The language of the content is the language in which the content is written.</td>
</tr>
<tr>
<td>License</td>
<td>The license of the content. The license is a legal instrument that grants permission or freedom to use, reproduce, and/or distribute the content.</td>
</tr>
<tr>
<td>SourceLocation</td>
<td>The location of the source. The location of the source where the content can be found.</td>
</tr>
<tr>
<td>SourceCountry</td>
<td>The country of the source. The country where the source of the content is located.</td>
</tr>
<tr>
<td>SourceCountryIso</td>
<td>The country ISO code of the source. The ISO code of the country where the source of the content is located.</td>
</tr>
</tbody>
</table>
Example I

• Inherited properties (from Creative Work and Thing)

Vocabulary – schema.org

• Example with microdata*:

```html
<div itemscope itemtype="http://schema.org/Movie">
<h1 itemprop="name">Avatar</h1>
<div itemprop="director" itemscope itemtype="http://schema.org/Person">
  Director:
  <span itemprop="name">James Cameron</span>
  (born
  <span itemprop="birthDate">August 16, 1954</span>)
</div>
<span itemprop="genre">Science fiction</span>
<a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

* http://schema.org/docs/gs.html
Other related vocabularies

- Can be mapped to other vocabularies such as DBPedia:
  - [http://dbpedia.org/ontology/](http://dbpedia.org/ontology/)
  - Link by using e.g. owl:equivalentProperty

Related Resources

- Web Data Commons

- Web Data Commons microdata corpus provides class-specific subsets of schema.org annotations that can be directly used as the working dataset

- The subsets contain all instances of a specific class of schema.org as well as all other data that is found on the webpages containing these instances.

Related Resources

- TopBraidComposer
  - Schema.org vocabularies already included

- GetSchema.org

- Schema 101: how to implement schema.org

Structured Data Testing Tool

- Test if the rich snippets are properly configured

- http://www.google.com/webmasters/tools/richsnippets
Structured Data Testing Tool

• Example: https://www.innsbruck.info/unterkuenfte/detail/unterkunft/grand-hotel-europa-innsbruck.html

Structured Data Testing Tool (New)

• https://developers.google.com/webmasters/structured-data/testing-tool/
Structured Data Marker Helper

- Assistant to annotate content with schema.org
- http://www.google.com/webmasters/tools/richsnippets

Schema Creator

- Provides templates to create annotations with schema.org and microdata for the most common vocabularies: Person, Product, Event, Organization, Movie, Book and Review.
- http://schema-creator.org/
Schema Creator - WordPress


- Schema Creator by Raven WordPress plugin simplifies the process of adding schema.org structured data to content published with WordPress.

- Provides an easy to use form to embed properly constructed schema.org microdata into a Wordpress post or page

Example of Schema.org Use: TVB Innsbruck Case

- Collaboration started in 2013 (STI & TVB Innsbruck)
- Strategies to enhance the visibility of their website and deal with the multi-channel communication challenges.
  - Semantic annotation in the website, blog
  - Dissemination of content with ONLIM
The Solution: implementation

http://blog.innsbruck.info/en

Schema.org for

✓ Restaurant, Cafes, Bars & Pubs, Sightseeing
  • Name
  • Map
  • PostalAddress
    o streetAddress
    o addressCountry
    o postalCode
    o addressLocality
    o telephone
    o faxNumber
Example of Café-Restaurant Villa Blanka

Object type: http://schema.org/Restaurant
Name: Café-Restaurant Villa Blanka
Address: http://schema.org/PostalAddress
  Street address: Weiherburggasse 8
  Address country: AT
  Postal code: 6020
  Address locality: innsbruck
  Telephone: +43 512 27 60 70

Implementation of semantic annotation with a plugin (Feratel -> Typo3)
ILLUSTRATION BY A LARGE EXAMPLE
Step 1: Opening the document

Open the document or write in the URL:

![Image of GATE software interface]

Step 2: Creating the Pipeline

Create pipeline for NLP processing by choosing the NLP applications, giving in the resources you want to process and appropriate parameters for them, then run this application:

![Image of GATE software interface with pipeline settings]
Step 3: Proving the automatic annotations

Prove the annotations made automatically and add your changes:

Step 4: Correcting the automated annotations:

Click on the items you want to change with the right mouse button and then change the annotation, add new annotation, or remove the existing annotation:
Choose from the tags offered or write in your annotation

Search for the entries of the expression in the whole text and annotate them

Remove annotation

Change the length of annotation

Step 5: Done!

Annotation after implementation of NLP techniques:

The Semantic Technology Institute (STI) Innsbruck, formerly known as DERI Innsbruck, was founded by Univ.-Prof. Dr. Dieter Fensel in 2002 and has developed into a challenging and dynamic research institute of approximately 80 people. STI Innsbruck collaborates with an international network of institutes in Asia, Europe and the USA, as well as with global industrial partners.

Final, manually-proved annotation:

The Semantic Technology Institute (STI) Innsbruck, formerly known as DERI Innsbruck, was founded by Univ.-Prof. Dr. Dieter Fensel in 2002 and has developed into a challenging and dynamic research institute of approximately 80 people. STI Innsbruck collaborates with an international network of institutes in Asia, Europe and the USA, as well as with global industrial partners.
SUMMARY

Summary (1)

• The population of ontologies is a task within the semantic content creation process as it links abstract knowledge to concrete knowledge.

• This knowledge acquisition can be done manually, semi-automatically, or fully automatically.

• There is a wide range of approaches that carry out semi-automatic annotation of text: most of the approaches make use of natural language processing and information extraction technology.

• In the annotation of multimedia aim at closing the so-called semantic gap, i.e. the discrepancy between low-level technical features which can be automatically processed to a large extent, and the high-level meaning-bearing features a user is typically interested in.

• Low level semantics can be extracted automatically, while high level semantics are still a challenge (and require human input to a large extent).
Summary (2)

- **Schema.org** provides a collection of shared vocabularies.

- Webmasters can use schema.org to mark up their web pages (creating enriched snippets) in a way that is recognized by major search engines.

- Search engines including Bing, Google, Yahoo! and Yandex rely on this markup to improve the display of search results.

- Most popular vocabularies related to Person, Place, LocalBusiness, Creative Work and Events.

- Schema.org can be used to enrich the web sites with the following formats: RDFa, microdata and JSON-LD.

REFERENCES
References

• Mandatory Reading:

• Further Reading:
  – GATE: http://gate.ac.uk/overview.html
  – Video Image Annotation Tool (formerly, M-OntoMat-Annotizer): https://sourceforge.net/projects/via-tool/
  – KIM platform (commercial product based on it): http://ontotext.com/semantic-solutions/dynamic-semantic-publishing-platform/
  – ALIPR: http://wang.ist.psu.edu/alipr/

References

– Games with a purpose: http://www.gwap.com
References

• Information for schema.org is taken from:
  – http://schema.org/docs/gs.html
  – http://moz.com/learn/seo/schema-structured-data

• Presentation of TVB Innsbruck use case by Renate Leitner and Anna Fensel, video at “Tourism Fast Forward” YouTube channel: https://www.youtube.com/watch?v=Vio8p4XIvRM (2014, ca. 45 minutes)

References

• Wikipedia links:
  – http://en.wikipedia.org/wiki/Games_with_a_purpose
### Next Lecture

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>Semantic Web Architecture</td>
</tr>
<tr>
<td>3</td>
<td>Resource Description Framework (RDF)</td>
</tr>
<tr>
<td>4</td>
<td>Web of data</td>
</tr>
<tr>
<td>5</td>
<td>Generating Semantic Annotations</td>
</tr>
<tr>
<td>6</td>
<td><strong>Storage and Querying</strong></td>
</tr>
<tr>
<td>7</td>
<td>Web Ontology Language (OWL)</td>
</tr>
<tr>
<td>8</td>
<td>Rule Interchange Format (RIF)</td>
</tr>
<tr>
<td>9</td>
<td>Reasoning on the Web</td>
</tr>
<tr>
<td>10</td>
<td>Ontologies</td>
</tr>
<tr>
<td>11</td>
<td>Social Semantic Web</td>
</tr>
<tr>
<td>12</td>
<td>Semantic Web Services</td>
</tr>
<tr>
<td>13</td>
<td>Tools</td>
</tr>
<tr>
<td>14</td>
<td>Applications</td>
</tr>
</tbody>
</table>

### Questions?