Semantic Web Technologies
Lecture 10: SWT for the Life Sciences 3: Text processing and ontologies

Maria Keet
email: keet -AT- inf.unibz.it
home: http://www.meteck.org
blog:
http://keet.wordpress.com/category/computer-science/72010-semwebtech/

KRDB Research Centre
Free University of Bozen-Bolzano, Italy

22 December 2009
Outline

Introduction

Ontology learning
  Background and methods
  Results and discussion

Ontology population
  Requirements for ontologies supporting NLP
  Results and discussion
Outline

Introduction

Ontology learning
  Background and methods
  Results and discussion

Ontology population
  Requirements for ontologies supporting NLP
  Results and discussion
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)
- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning (today; ref Alexopoulou et al., 2008)
- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al., 2007)
- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)
- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning (today; ref Alexopoulou et al., 2008)
- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al., 2007)
- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
    - To sort literature results
    - To navigate literature (linked data)
- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning (today; ref Alexopoulou et al., 2008)
- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al., 2007)
- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)
- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning (today; ref. Alexopoulou et al., 2008)
- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
    - Biomedical text mining (today; ref. Witt et al., 2007)
- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)

- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning
    (today; ref Alexopoulou et al, 2008)

- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al, 2007)

- Natural language generation from a formal language
Natural language and ontologies

- **Using ontologies to improve NLP**
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)
- **Using NLP to develop ontologies (TBox)**
  - Searching for candidate terms and relations: Ontology learning (today; ref Alexopoulou et al, 2008)
- **Using NLP to populate ontologies (ABox)**
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al, 2007)
- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)
- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning (today; ref Alexopoulou et al, 2008)
- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al, 2007)
- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)
- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning
    (today; ref Alexopoulou et al, 2008)
- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al, 2007)
  - Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)

- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning
    (today; ref Alexopoulou et al, 2008)

- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
    - Biomedical text mining (today; ref Witte et al, 2007)

- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)

- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning
    (today; ref Alexopoulou et al, 2008)

- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al, 2007)

- Natural language generation from a formal language
Natural language and ontologies

- Using ontologies to improve NLP
  - To enhance precision and recall of queries
  - To enhance dialogue systems
  - To sort literature results
  - To navigate literature (linked data)

- Using NLP to develop ontologies (TBox)
  - Searching for candidate terms and relations: Ontology learning
    (today; ref Alexopoulou et al., 2008)

- Using NLP to populate ontologies (ABox)
  - Document retrieval enhanced by lexicalised ontologies
  - Biomedical text mining (today; ref Witte et al., 2007)

- Natural language generation from a formal language
Semantic Tagging—Classes, Terms

Semantic Tagging—Lexicalized Ontologies

http://olp.dfki.de/LingInfo/
http://ontoware.org/projects/lexonto/

Examples (out of many)


- **GoPubMed** (Dietze et al, 2009)
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology

- **Question answer system AliQAn for agriculture** (Vila and Ferrández, 2009)
  - Question assignment task too difficult for specialised domains
  - Add ontology to an open domain QA system, using AGROVOC and WordNet

- **Attempto Controlled English (ACE), rabbit, etc.**; grammar engine, template-based approach
Examples (out of many)

- **GoPubMed** *(Dietze et al, 2009)*
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology
- **Question answer system AliQAn** for agriculture *(Vila and Ferrández, 2009)*
  - Question assignment task too difficult for specialised domains
  - Add ontology to an open domain QA system, using AGROVOC and WordNet
- **Attempto Controlled English (ACE)**, rabbit, etc.; grammar engine, template-based approach
Examples (out of many)


- **GoPubMed** (Dietze et al, 2009)
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology

- **Question answer system AliQAn for agriculture** (Vila and Ferrández, 2009)
  - Question assignment task too difficult for specialised domains
  - Add ontology to an open domain QA system, using AGROVOC and WordNet
  - Attempto Controlled English (ACE), rabbit, etc.; grammar engine, template-based approach
Examples (out of many)


- **GoPubMed** *(Dietze et al, 2009)*
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology

- **Question answer system AliQAn for agriculture** *(Vila and Ferrández, 2009)*
  - Question assignment task too difficult for specialised domains
  - Add ontology to an open domain QA system, using AGROVOC and WordNet

- **Attempto Controlled English (ACE), rabbit, etc.**; grammar engine, template-based approach
Examples (out of many)


- **GoPubMed** *(Dietze et al, 2009)*
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology

- **Question answer system AliQAn for agriculture** *(Vila and Ferrández, 2009)*
  - Question assignment task too difficult for specialised domains
  - Add ontology to an open domain QA system, using AGROVOC and WordNet
  - Attempto Controlled English (ACE), rabbit, etc.; grammar engine, template-based approach
Examples (out of many)


- **GoPubMed** (Dietze et al, 2009)
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology

- **Question answer system AliQAn for agriculture** (Vila and Ferrández, 2009)
  - Question assignment task too difficult for specialised domains
    - Add ontology to an open domain QA system, using AGROVOC and WordNet
  - Attempto Controlled English (ACE), rabbit, etc.; grammar engine, template-based approach
Examples (out of many)


- **GoPubMed** *(Dietze et al, 2009)*
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology

- **Question answer system AliQAn for agriculture** *(Vila and Ferrández, 2009)*
  - Question assignment task too difficult for specialised domains
  - Add ontology to an open domain QA system, using AGROVOC and WordNet

- Attempto Controlled English (ACE), rabbit, etc.; grammar engine, template-based approach
Examples (out of many)


- **GoPubMed** (Dietze et al, 2009)
  - Layer over PubMed, which indexes ± 19mln articles in the bio(medical) domain; pre-processing of the abstracts (advanced semantic tagging)
  - Results of the PubMed query are sorted according to terms in the ontology

- **Question answer system AliQAn for agriculture** (Vila and Ferrández, 2009)
  - Question assignment task too difficult for specialised domains
  - Add ontology to an open domain QA system, using AGROVOC and WordNet

- **Attempto Controlled English (ACE), rabbit, etc.**; grammar engine, template-based approach
Outline

Introduction

Ontology learning
  Background and methods
  Results and discussion

Ontology population
  Requirements for ontologies supporting NLP
  Results and discussion
Outline

Introduction

Ontology learning
  Background and methods
  Results and discussion

Ontology population
  Requirements for ontologies supporting NLP
  Results and discussion
Background

- Ontology development is time consuming
- Bottom-up ontology development strategies discussed in lecture 4, of which one is to use NLP
- Where, if anywhere, can NLP make life easier for ontology development, and how?
- Current results are mostly discouraging, and depend on the approach, technique, and ontological commitment
  - We take a closer look at ontology learning limited to finding terms for a domain ontology
Background

- Ontology development is time consuming
- Bottom-up ontology development strategies discussed in lecture 4, of which one is to use NLP
- Where, if anywhere, can NLP make life easier for ontology development, and how?
- Current results are mostly discouraging, and depend on the approach, technique, and ontological commitment
  - We take a closer look at ontology learning limited to finding terms for a domain ontology
Background

- Ontology development is time consuming
- Bottom-up ontology development strategies discussed in lecture 4, of which one is to use NLP
- Where, if anywhere, can NLP make life easier for ontology development, and how?
- Current results are mostly discouraging, and depend on the approach, technique, and ontological commitment
  - We take a closer look at ontology learning limited to finding terms for a domain ontology
Bottom-up ontology development with NLP

• Usual parameters, such as purpose (in casu, document retrieval), formal language (an OWL species)
• A standard kind of ontology (not a comprehensive lexicalised ontology)
• Additional considerations for “text-mining ontologies”
  • Level of granularity of the terms to include (hypo/hypernyms)
  • How to deal with synonyms (‘LDL I’ and ‘large LDL’)
  • Handle term variations (e.g., ‘LDL-I’ and ‘LDL I’, ‘Tangiers’ disease’ and ‘Tangier’s Disease’)
  • Disambiguation; e.g. w.r.t. abbreviations
Method to test automated term recognition

- Compare the terms of a manually constructed ontology with the terms obtained from text mining a suitable corpus
- Build an ontology manually
  - Lipoprotein metabolism (LMO), 223 classes with 623 synonyms
- Create a corpus
  - 3066 review article abstract from PubMed, obtained with a ‘lipoprotein metabolism’ search
- Automatic Term Recognition (ATR) tools
  - Text2Onto: relative term frequency, TFIDF, entropy, hypernym structure of WordNet, Hearst patterns
  - Termine: statistics of candidate term, such as total frequency of occurrence, frequency of the term as part of other longer candidate terms, length of term
  - OntoLearn: linguistic processor and syntactic parser, Domain relevance and domain consensus
  - RelFreq: relative frequency of a term in a corpus
  - TFIDF: RelFreq + doc. frequency derived from all phrases in PubMed
Outline

Introduction

Ontology learning
  Background and methods
  Results and discussion

Ontology population
  Requirements for ontologies supporting NLP
  Results and discussion
Results

- OntoLearn excluded form analysis because it regenerated few terms
- Text2Onto only included in analysis for up to 300 abstracts (could not process all 3066)
- Precision for LMO 17-35% for top 50 terms, and 4-8% for top 1000 terms
- Precision for LMO + expert analysis of the automatically generated terms: up to 75% for top 50 terms, and up to 29% for top 1000 terms
- Termine good for the longer terms, RelFreq and TFIDF for the shorter terms
Results (cont’d)

Table 3: Coverage of LMO terminology in selected document sets. The table sets the upper limit of terms that can be found with text-mining. Even a large text base with 50,000 documents contains only 71% of LMO terms. TFIDF can predict up to 38% of LMO terms.

<table>
<thead>
<tr>
<th>LMO terminology predicted by TFIDF</th>
<th>LMO terminology literally contained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000</td>
</tr>
<tr>
<td>300 review abstracts for “lipoprotein metabolism”</td>
<td>8.75%</td>
</tr>
<tr>
<td>3,066 abstracts for “lipoprotein metabolism”</td>
<td>14.99%</td>
</tr>
<tr>
<td>50,000 abstracts containing “lipoprotein”</td>
<td></td>
</tr>
</tbody>
</table>

from Alexopoulou et al, 2008
What went wrong with some of the terms?

- LMO terms that were not in the 50k abstracts grouped into:
  - Rarely occurring terms: occur rarely even in the whole of PubMed
  - Rarely occurring variants of terms: e.g., ‘free chol’ (0, instead of 2622 for ‘free cholesterol’)
  - Very long terms; e.g., ‘predominance of large low-density lipoprotein particles’, which can be decomposed into smaller terms
  - Combinations of terms/variants; e.g., ‘increased total chol’ (0, instead of 116 for ‘increased total cholesterol’),
  - Terms that should normally be easily found; e.g., ‘diabetes type I’ (126) and ‘acetyl-coa c-acyltransferase’, probably due to limited corpus
  - Predicted terms, not in LMO: wrongly predicted (±25% of the TFIDF top50) or can be added to LMO (±40% of the TFIDF top50)
What went wrong with some of the terms?

- LMO terms that were not in the 50k abstracts grouped into:
  - Rarely occurring terms: occur rarely even in the whole of PubMed
  - Rarely occurring variants of terms: e.g., ‘free chol’ (0, instead of 2622 for ‘free cholesterol’)
  - Very long terms; e.g., ‘predominance of large low-density lipoprotein particles’, which can be decomposed into smaller terms
  - Combinations of terms/variants; e.g., ‘increased total chol’ (0, instead of 116 for ‘increased total cholesterol’),
  - Terms that should normally be easily found; e.g., ‘diabetes type I’ (126) and ‘acetyl-coa c-acyltransferase’, probably due to limited corpus

- Predicted terms, not in LMO: wrongly predicted (±25% of the TFIDF top50) or can be added to LMO (±40% of the TFIDF top50)
What went wrong with some of the terms?

- LMO terms that were not in the 50k abstracts grouped into:
  - Rarely occurring terms: occur rarely even in the whole of PubMed
  - Rarely occurring variants of terms: e.g., ‘free chol’ (0, instead of 2622 for ‘free cholesterol’)  
  - Very long terms; e.g., ‘predominance of large low-density lipoprotein particles’, which can be decomposed into smaller terms
  - Combinations of terms/variants; e.g., ‘increased total chol’ (0, instead of 116 for ‘increased total cholesterol’),
  - Terms that should normally be easily found; e.g., ‘diabetes type I’ (126) and ‘acetyl-coa c-acyltransferase’, probably due to limited corpus

- Predicted terms, not in LMO: wrongly predicted (±25% of the TFIDF top50) or can be added to LMO (±40% of the TFIDF top50)
What went wrong with some of the terms?

- LMO terms that were not in the 50k abstracts grouped into:
  - Rarely occurring terms: occur rarely even in the whole of PubMed
  - Rarely occurring variants of terms: e.g., ‘free chol’ (0, instead of 2622 for ‘free cholesterol’)
  - Very long terms; e.g, ‘predominance of large low-density lipoprotein particles’, which can be decomposed into smaller terms
  - Combinations of terms/variants; e.g., ‘increased total chol’ (0, instead of 116 for ‘increased total cholesterol’),
  - Terms that should normally be easily found; e.g., ‘diabetes type I’ (126) and ‘acetyl-coa c-acyltransferase’, probably due to limited corpus
  - Predicted terms, not in LMO: wrongly predicted (±25% of the TFIDF top50) or can be added to LMO (±40% of the TFIDF top50)
What went wrong with some of the terms?

- LMO terms that were not in the 50k abstracts grouped into:
  - Rarely occurring terms: occur rarely even in the whole of PubMed
  - Rarely occurring variants of terms: e.g., ‘free chol’ (0, instead of 2622 for ‘free cholesterol’)
  - Very long terms; e.g, ‘predominance of large low-density lipoprotein particles’, which can be decomposed into smaller terms
  - Combinations of terms/variants; e.g., ‘increased total chol’ (0, instead of 116 for ‘increased total cholesterol’),
  - Terms that should normally be easily found; e.g., ‘diabetes type I’ (126) and ‘acetyl-coa c-acyltransferase’, probably due to limited corpus

- Predicted terms, not in LMO: wrongly predicted ($\pm 25\%$ of the TFIDF top50) or can be added to LMO ($\pm 40\%$ of the TFIDF top50)
What went wrong with some of the terms?

- LMO terms that were not in the 50k abstracts grouped into:
  - Rarely occurring terms: occur rarely even in the whole of PubMed
  - Rarely occurring variants of terms: e.g., ‘free chol’ (0, instead of 2622 for ‘free cholesterol’)
  - Very long terms; e.g., ‘predominance of large low-density lipoprotein particles’, which can be decomposed into smaller terms
  - Combinations of terms/variants; e.g., ‘increased total chol’ (0, instead of 116 for ‘increased total cholesterol’),
  - Terms that should normally be easily found; e.g., ‘diabetes type I’ (126) and ‘acetyl-coa c-acyltransferase’, probably due to limited corpus

- Predicted terms, not in LMO: wrongly predicted (±25% of the TFIDF top50) or can be added to LMO (±40% of the TFIDF top50)
What went wrong with some of the terms?

- LMO terms that were not in the 50k abstracts grouped into:
  - Rarely occurring terms: occur rarely even in the whole of PubMed
  - Rarely occurring variants of terms: e.g., ‘free chol’ (0, instead of 2622 for ‘free cholesterol’)
  - Very long terms; e.g., ‘predominance of large low-density lipoprotein particles’, which can be decomposed into smaller terms
  - Combinations of terms/variants; e.g., ‘increased total chol’ (0, instead of 116 for ‘increased total cholesterol’),
  - Terms that should normally be easily found; e.g., ‘diabetes type I’ (126) and ‘acetyl-coa c-acyltransferase’, probably due to limited corpus

- Predicted terms, not in LMO: wrongly predicted (±25% of the TFIDF top50) or can be added to LMO (±40% of the TFIDF top50)
### Outline

#### Introduction

#### Ontology learning
- Background and methods
- Results and discussion

#### Ontology population
- Requirements for ontologies supporting NLP
- Results and discussion
Outline

Introduction

Ontology learning
  Background and methods
  Results and discussion

Ontology population
  Requirements for ontologies supporting NLP
  Results and discussion

Summary
Typical NLP tasks

- **Named Entity recognition/semantic tagging;** e.g., “... the organisms were incubated at 37°C”
- **Entity normalization;** e.g., different strings refer to the same thing (full and abbreviated name, or single letter amino acid, three-letter aminoacid and full name: W, Trp, Tryptophan)
- **Coreference resolution;** in addition to synonyms (lactase and β-galactosidase), there as pronominal references (it, this)
- **Grounding;** the text string w.r.t. external source, like UniProt, that has the representation of the entity in reality
- **Relation detection;** most of the important information in contained within the relations between entities, NLP can be enhanced by considering semantically possible relations
Typical NLP tasks

- Named Entity recognition/semantic tagging; e.g., “… the organisms were incubated at 37°C”
- Entity normalization; e.g., different strings refer to the same thing (full and abbreviated name, or single letter amino acid, three-letter aminoacid and full name: W, Trp, Tryptophan)
- Coreference resolution; in addition to synonyms (lactase and β-galactosidase), there are pronominal references (it, this)
- Grounding; the text string w.r.t. external source, like UniProt, that has the representation of the entity in reality
- Relation detection; most of the important information is contained within the relations between entities, NLP can be enhanced by considering semantically possible relations
Typical NLP tasks

- Named Entity recognition/semantic tagging; e.g., “... the organisms were incubated at 37°C”
- Entity normalization; e.g., different strings refer to the same thing (full and abbreviated name, or single letter amino acid, three-letter aminoacid and full name: W, Trp, Tryptophan)
- Coreference resolution; in addition to synonyms (lactase and β-galactosidase), there as pronominal references (it, this)
- Grounding; the text string w.r.t. external source, like UniProt, that has the representation of the entity in reality
- Relation detection; most of the important information in contained within the relations between entities, NLP can be enhanced by considering semantically possible relations
Typical NLP tasks

- Named Entity recognition/semantic tagging; e.g., “... the organisms were incubated at 37°C”
- Entity normalization; e.g., different strings refer to the same thing (full and abbreviated name, or single letter amino acid, three-letter amino acid and full name: W, Trp, Tryptophan)
- Coreference resolution; in addition to synonyms (lactase and β-galactosidase), there are pronominal references (it, this)
- Grounding; the text string w.r.t. external source, like UniProt, that has the representation of the entity in reality
- Relation detection; most of the important information is contained within the relations between entities, NLP can be enhanced by considering semantically possible relations
Typical NLP tasks

- Named Entity recognition/semantic tagging; e.g., “… the organisms were incubated at 37°C”
- Entity normalization; e.g., different strings refer to the same thing (full and abbreviated name, or single letter amino acid, three-letter amino acid and full name: W, Trp, Tryptophan)
- Coreference resolution; in addition to synonyms (lactase and β-galactosidase), there as pronominal references (it, this)
- Grounding; the text string w.r.t. external source, like UniProt, that has the representation of the entity in reality
- Relation detection; most of the important information in contained within the relations between entities, NLP can be enhanced by considering semantically possible relations
Requirements for NLP ontologies

- Domain ontology (at least a taxonomy)
  - Text model, concerns with classes such as sentence, text position and locations like abstract, introduction
  - Biological entities, i.e., contents for the ABox, often already available in biological databases on the Internet
  - Lexical information for recognizing named entities; full names of entities, their synonyms, common variants and misspellings, and knowledge about naming, like endo- and -ase
  - Database links to connect the lexical term to the entity represent in a particular database (the grounding step)
  - Entity relations; represented in the domain ontology
Requirements for NLP ontologies

- Domain ontology (at least a taxonomy)
- Text model, concerns with classes such as sentence, text position and locations like abstract, introduction
- Biological entities, i.e., contents for the ABox, often already available in biological databases on the Internet
- Lexical information for recognizing named entities; full names of entities, their synonyms, common variants and misspellings, and knowledge about naming, like endo- and -ase
- Database links to connect the lexical term to the entity represent in a particular database (the grounding step)
- Entity relations; represented in the domain ontology
Requirements for NLP ontologies

- Domain ontology (at least a taxonomy)
- Text model, concerns with classes such as sentence, text position and locations like abstract, introduction
- Biological entities, i.e., contents for the ABox, often already available in biological databases on the Internet
- Lexical information for recognizing named entities; full names of entities, their synonyms, common variants and misspellings, and knowledge about naming, like endo- and -ase
- Database links to connect the lexical term to the entity represent in a particular database (the grounding step)
- Entity relations; represented in the domain ontology
Requirements for NLP ontologies

- Domain ontology (at least a taxonomy)
- Text model, concerns with classes such as sentence, text position and locations like abstract, introduction
- Biological entities, i.e., contents for the ABox, often already available in biological databases on the Internet
- Lexical information for recognizing named entities; full names of entities, their synonyms, common variants and misspellings, and knowledge about naming, like endo- and -ase
  - Database links to connect the lexical term to the entity represent in a particular database (the grounding step)
  - Entity relations; represented in the domain ontology
Requirements for NLP ontologies

• Domain ontology (at least a taxonomy)
• Text model, concerns with classes such as sentence, text position and locations like abstract, introduction
• Biological entities, i.e., contents for the ABox, often already available in biological databases on the Internet
• Lexical information for recognizing named entities; full names of entities, their synonyms, common variants and misspellings, and knowledge about naming, like endo- and -ase
• Database links to connect the lexical term to the entity represent in a particular database (the grounding step)
• Entity relations; represented in the domain ontology
Requirements for NLP ontologies

- Domain ontology (at least a taxonomy)
- Text model, concerns with classes such as *sentence*, *text position* and locations like *abstract*, *intorduction*
- Biological entities, i.e., contents for the ABox, often already available in biological databases on the Internet
- Lexical information for recognizing named entities; full names of entities, their synonyms, common variants and misspellings, and knowledge about naming, like *endo-* and *-ase*
- Database links to connect the lexical term to the entity represent in a particular database (the grounding step)
- Entity relations; represented in the domain ontology
Outline

Introduction

Ontology learning
   Background and methods
   Results and discussion

Ontology population
   Requirements for ontologies supporting NLP
   Results and discussion
MutationMiner use case

• See Witte et al. book chapter for details
  • Ontology in OWL, in Protégé; with class name, textual definition and example instances
  • Species info from the NCBI taxonomy; note the management of central \textit{scientific name} and its synonyms, common variants and misspellings
  • Uniprot and use of its back-links to the NCBI taxonomy
MutationMiner use case

- See Witte et al. book chapter for details
- Ontology in OWL, in Protégé; with class name, textual definition and example instances
  - Species info from the NCBI taxonomy; note the management of central *scientific name* and its synonyms, common variants and misspellings
  - Uniprot and use of its back-links to the NCBI taxonomy
MutationMiner use case

- See Witte et al. book chapter for details
- Ontology in OWL, in Protégé; with class name, textual definition and example instances
- Species info from the NCBI taxonomy; note the management of central *scientific name* and its synonyms, common variants and misspellings
- Uniprot and use of its back-links to the NCBI taxonomy
MutationMiner use case

- See Witte et al. book chapter for details
- Ontology in OWL, in Protégé; with class name, textual definition and example instances
- Species info from the NCBI taxonomy; note the management of central scientific name and its synonyms, common variants and misspellings
- Uniprot and use of its back-links to the NCBI taxonomy
Introduction

Ontology learning

Ontology population

Summary

Discussion

• See Witte et al. book chapter for details

• Significant upfront investments due to novelty and complexity of SWT

• Benefits:
  • Standardizes data exchange, consolidate disparate resources
  • Detecting inconsistencies (caused by, e.g., a pronoun with an incompatible relation to another textual entity)

• To do: Ontological NLP, enhancing standard NLP tools to take more of SWT into account
Discussion

• See Witte et al. book chapter for details
• Significant upfront investments due to novelty and complexity of SWT
• Benefits:
  • Standardizes data exchange, consolidate disparate resources
  • Detecting inconsistencies (caused by, e.g., a pronoun with an incompatible relation to another textual entity)
• To do: Ontological NLP, enhancing standard NLP tools to take more of SWT into account
Discussion

• See Witte et al. book chapter for details
• Significant upfront investments due to novelty and complexity of SWT
• Benefits:
  • Standardizes data exchange, consolidate disparate resources
  • Detecting inconsistencies (caused by, e.g. a pronoun with an incompatible relation to another textual entity)
• To do: Ontological NLP, enhancing standard NLP tools to take more of SWT into account
Discussion

- See Witte et al. book chapter for details
- Significant upfront investments due to novelty and complexity of SWT
- Benefits:
  - Standardizes data exchange, consolidate disparate resources
  - Detecting inconsistencies (caused by, e.g. a pronoun with an incompatible relation to another textual entity)
- To do: Ontological NLP, enhancing standard NLP tools to take more of SWT into account
Discussion

• See Witte et al. book chapter for details
• Significant upfront investments due to novelty and complexity of SWT
• Benefits:
  • Standardizes data exchange, consolidate disparate resources
  • Detecting inconsistencies (caused by, e.g. a pronoun with an incompatible relation to another textual entity)
• To do: Ontological NLP, enhancing standard NLP tools to take more of SWT into account
Discussion

- See Witte et al. book chapter for details
- Significant upfront investments due to novelty and complexity of SWT
- Benefits:
  - Standardizes data exchange, consolidate disparate resources
  - Detecting inconsistencies (caused by, e.g. a pronoun with an incompatible relation to another textual entity)
- To do: Ontological NLP, enhancing standard NLP tools to take more of SWT into account
Summary

Introduction

Ontology learning
  Background and methods
  Results and discussion

Ontology population
  Requirements for ontologies supporting NLP
  Results and discussion