Outline

Part 1: Ingredients and Techniques

Part 2: Recipes
### Part 1: Ingredients and Techniques

**Ingredients:** Lexical Resources

**Integration Technique:** Automatic Sense Linking

**Interoperability Technique:** Standardizing
Part 1: Ingredients and Techniques

Ingredients: Lexical Resources

Elements of Lexical Resources

Classic Lexical Resources

Collaborative Lexical Resources

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing
Elements of Lexical Resources

Lexical Markup Framework (LMF, ISO 24613:2008) – Core Package

- A lexical resource consists of lexicons.
- A lexicon belongs to a particular language and consists of lexical entries.
- LexicalEntry is a class representing a lexeme in a given language.
- A lexeme is an abstract pairing of meaning and form (Jurafsky & Martin, 2008)
The Form Part of a Lexical Entry: Lemma and Word Form

Lexical entry
POS: verb

Lemma
sing

Word Form
sang

Word Form
sung
The Meaning Part of a Lexical Entry: Sense

- Lemmas can have several senses (*lexical ambiguity*)
- Colloquial: “*words can have several meanings*”
- Many lemmas are associated with more than one sense

```
Lemma
to sing
```

<table>
<thead>
<tr>
<th>Sense #1</th>
<th>Sense #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sing / produce tones</td>
<td>Sing / divulge secrets</td>
</tr>
</tbody>
</table>

1. Produce tones with the voice
2. divulge confidential information or secrets
Lexical Information about Senses I

- **Definitions**: short summaries of the meaning of a sense, also called *glosses*. They are meant to define a meaning.

  - The lemma *sing* has a sense that can be defined as “*Produce tones with the voice*“.

- **Sense examples**: Senses can also be described by example sentences. They illustrate a meaning.

  - Sense example for lemma *sing* in the sense “*Produce tones with the voice*“: "*She was singing while she was cooking.*"
Lexical Information about Senses II

- Translations
  - German translation of *to sing*: *singen*

- Usage information, e.g. register (informal, slang, …)
  - *sing / divulge secrets* has register *slang*

- Semantic field
  - e.g., in WordNet: animate, food, location, communication …
  - semantic field of *sing / produce tones*: *creation*

- Domain:
  - *sing / produce tones* has domain *music*
Lexical Information about Senses III

Morphologically Related Senses

- E.g. *sing / produce tones with the voice* (verb) – *singer* (noun)

Sense Relations

- Synonymy: equivalent senses are related by the synonymy relation
  - Colloquial: „*the same meaning can be expressed by different words*“

- Hypernymy / Hyponymy (noun senses): Also called the IS-A relation
  - *singer* is-a *musician*

- Many more sense relations …
Lexical Information about Senses IV

- **Syntactic behavior**: Subcategorization frames (SCFs)
- SCFs specify
  - syntactic categories (NP-nominative, NP-accusative, PP …) and
  - grammatical functions (subject, object, ...) of the arguments of a verbal, nominal or adjectival predicate.
- partly language-specific

```
She                          is singing.
[subject, nominative]

She                          is singing Christmas carols.
[subject, nominative]        [object, accusative]
```
**Relation of Sense and Subcategorization Frame**

Example:

- *Sing / produce tones with the voice* can be used with NP-accusative:
  - They sing *Christmas carols*.

- *Sing / divulge secrets* can not be used with NP-accusative:
  - Mob Informant Joe Vollaro sings again.
Grouping Related Senses – Different Ways to Organize a Lexicon

In dictionaries and in the ISO standard LMF, senses are grouped into lexical entries that share the same lemma and part-of-speech.

However, senses can also be grouped differently:

- Grouping of senses that are related by some sense relation
  - e.g., synonymy

- Grouping of senses that share the same syntactic behavior
  - e.g., subcategorization frame
### Part 1: Ingredients and Techniques

**Ingredients: Lexical Resources**

<table>
<thead>
<tr>
<th>Elements of Lexical Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic Lexical Resources</td>
</tr>
<tr>
<td>Collaborative Lexical Resources</td>
</tr>
</tbody>
</table>

**Integration Technique: Automatic Sense Linking**

**Interoperability Technique: Standardizing**
WordNet

- Domain-independent, broad-coverage lexical-semantic network of English nouns, verbs and adjectives

- Realized at Princeton University by George Miller’s team (started in 1985)

- Widely used for many NLP tasks and applications

- [http://wordnet.princeton.edu/wordnet/](http://wordnet.princeton.edu/wordnet/)
WordNet is Organized in Synsets

Synonymous senses are grouped into synsets.

spill the beans#1
let the cat out of the bag#1
talk#5	
tattle#2
blab#1
peach#1
babble#4
**sing**#5
babble out#1
blab out#1

**Synset**,
Synset ID: (1){00939238}
WordNet Synsets

Synonymous senses are grouped into synsets.

Members of a synset
- are senses,
- are represented by their lemma.

All synset members belong to the same word class (e.g., noun, verb).

spill the beans#1
let the cat out of the bag#1
talk#5
tattle#2
blab#1
peach#1
babble#4
\textbf{sing#5}
babble out#1
blab out#1

\textbf{Synset ,}
Synset ID: (1){00939238}
Synset Definitions (Glosses)

Synsets have glosses, i.e. definitions/short summaries of their meaning.

- The meaning of a synset is expressed in its gloss.
- The meaning of a synset can alternatively be captured by the list of its member synonyms.

**Synset**, Synset ID: (1){00939238}

**Gloss**: “divulge confidential information or secrets“

**Members:**
- *sing* (Sense ID: sing%2:32:01::)
- *spill the beans* (Sense ID: spill_the_beans%2:32:00::)
Sense Examples

Senses, i.e., members of a synset, are illustrated by sense examples – sentences illustrate a meaning.

**Synset**, Synset ID: (1){00939238}

**Gloss**: “divulge confidential information or secrets“

**Members:**

*Sing* (Sense ID: sing%2:32:01::)

**Sense Example:** "Mob Informant Joe Vollaro sings again"

*Spill the beans* (Sense ID: spill_the_beans%2:32:00::)

**Sense Example:** "They had planned it as a surprise party, but somebody spilled the beans."
VerbNet

- Domain-independent, broad-coverage lexicon of English verbs
  - lexical-syntactic and lexical-semantic information for verbs
  - 3962 verb lemmas in VerbNet 3.1

- [http://verbs.colorado.edu/~mpalmer/projects/verbnet.html](http://verbs.colorado.edu/~mpalmer/projects/verbnet.html)
VerbNet is Organized in Verb Classes

Verbs are grouped into **verb classes**, based on Levin’s classification of English verbs:

- 470 verb classes, hierarchically structured
- **VerbNet sense** = pair of verb lemma and verb class

chant chatter chirp chortle
chuckle cluck coo croak ... scream screech shout shriek sibilate
sigh simper

**sing** smatter smile snap
snarl snivel
snuffle ...

**manner_speaking-37.3**
Members: 103, Frames: 14
VerbNet Verb Classes

VerbNet verb classes group verbs that share the same predicate-argument structure, i.e.
- subcategorization frame,
- semantic roles and selectional preferences,
- semantic predicate based on the event decomposition of Moens and Steedman (1988).

Although the resulting verb classes are semantically coherent, the semantic relatedness of verb senses in a VerbNet class is **distant** compared to WordNet synsets.
- e.g., the verbs *believe*, *swear* and *doubt* are in the same verb class.
VerbNet – Information types

- subcategorization frame, semantic roles and selectional preferences
- semantic predicate
- single-verb sense example in the lexical resource VerbNet
- SemLink provides links to real sense examples in PropBank

Example:

**VERB:** sing

**EXAMPLE:** "Susan whispered about the party."

**SYNTAX:**
Agent[+animate|+organization] V {about} Topic[+communication]

**SEMANTIC PRED:**
transfer_info(during(E), Agent, ?Recipient, Topic)
cause(Agent, E)
FrameNet

FrameNet is a lexical resource of English verbs, nouns, adjectives.

FrameNet 1.5 contains:
- 1,019 frames,
- 9,423 lemmas,
- 11,942 lexical units

http://framenet.icsi.berkeley.edu/
FrameNet is Organized in Frames

- Senses are grouped into **frames** based on Fillmore’s Frame Semantics (Fillmore 1976, Fillmore et al. 2003).
- In FrameNet, senses are called lexical units.

**Communication_manner**

babble.v, bluster.v, chant.v, chatter.v, drawl.v, gabble.v, gibber.v, jabber.v, lisp.v, mouth.v, mumble.v, mutter.v, natter.v, prattle.v, rant.v, rave.v, shout.v, simper.v, **sing.v**, slur.v, stammer.v, stutter.v, whisper.v
FrameNet Frames

- A frame represents a conceptual structure, or a prototypical situation with a (frame-specific) set of roles that identify the participants involved in the situation.

- Frames group senses which evoke the same kind of situation with participants taking over particular roles.

- Senses in a FrameNet frame are semantically related, but not synonymous; e.g., the verbs love and hate are both in the same FrameNet frame.
Frame Evoking Word Classes

Frame evoking word classes are

- **verbs**, they are the prototypical frame-evoking word classes
- **predicate-like** nouns and adjectives
  - e.g. nouns denoting events (*development*), relations (*brother*), states (*height*)

Example:

COMMUNICATION_MANNER frame:

\[
[\text{They}]_{\text{Speaker}} \text{ all } \text{sang} [\text{Happy Birthday}]_{\text{Message}}
\]
Coverage Issues in FrameNet

Low lexical coverage compared to WordNet
- 9,423 lemmas in FrameNet 1.5
- 156,584 lemmas in WordNet 3.0

- The focus of FrameNet is on verbs and predicate-like nouns and adjectives.
  - Many nouns and adjectives evoke uninteresting frames, e.g., nouns denoting artifacts and natural kinds; adjectives denoting colors.
  - Hence, few of them have been included [Baker and Fellbaum 2009].

- Senses encountered in a corpus may have no corresponding sense in FrameNet, e.g.
  
  Sing / produce tones with the voice → Frame Communication_manner
  Sing / divulge secrets → No frame available
Part 1: Ingredients and Techniques

Ingredients: Lexical Resources

Elements of Lexical Resources

Classic Lexical Resources

Collaborative Lexical Resources

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing
Wikipedia, a Multilingual Encyclopedia

- Wikipedia is a freely licensed encyclopedia written by thousands of volunteers in many languages.

- Free license allows others to freely copy, redistribute, and modify the work commercially or non-commercially.

- Founded January 15, 2001

- 6th most visited site according to Alexa (Feb. 2010)

http://www.wikipedia.org

(Jimmy Wales)
Wikipedia – Information Types

**Title**: Benzoic acid

**Redirects**: From Wikipedia, the free encyclopedia (Redirected from Benzenecarboxylic acid)

**Introduction (first paragraph)**: Benzoic acid, C₆H₅CO₂H (or C₆H₅COOH), is a colorless crystalline solid and the simplest aromatic carboxylic acid. The name derived from gum benzoïn, which was for a long time the only source for benzoic acid. This weak acid and its salts are used as a food preservative. Benzoic acid is an important precursor for the synthesis of many other organic substances.

**Headers**

1. History
2. Production
   2.1 Industrial preparations
      2.2 Laboratory synthesis
         2.2.1 By hydrolysis
         2.2.2 From benzaldehyde
         2.2.3 From bromobenzene
         2.2.4 From benzyl alcohol
3. Uses
   3.1 Food
   3.2 Feedstock
   3.3 Medicinal
4. Biology and health effects
5. Chemistry

**Categories**

- Organic acids
- Benzoic acids
- Aromatic compounds
- Excipients

**Infoboxes**

**Hyperlinks**

- Benzoic acid: esters are the product of the acid–alcohol reaction with alcohols.
- Benzoic acid: amides are more easily available by using activated acid derivatives (such as benzoyl chloride) or by coupling reagents used in peptide synthesis like DCC and DMAP.

---

09.2013 | Computer Science Department | UKP Lab | Iryna Gurevych and Judith Eckle-Kohler
Wikipedia – Disambiguation pages

- Sense inventory, including domain specific senses

---

Forest (disambiguation)

From Wikipedia, the free encyclopedia

A forest is a large area covered by trees.

Forest can also mean:

- Royal forest, an area set aside for hunting

Forest may also be:

- In Windows networking, the collection of every object, their attributes and rules in an Active Directory
- In graph theory, a disjoint union of trees
- Forest (album), an album by George Winston
- "Forest" (song), a song by the band System of a Down

The Forest may refer to:

- The Forest, a video game
- The Forest, a 2002 film
JWPL
Java-based Wikipedia Programming Library

Highlights

- High performance access to Wikipedia content
- Parser for the WikiMedia syntax
- Articles, discussion pages, categories as Java objects
- Access to information nuggets
  - Redirects, links, link anchors, interlanguage links, sections, first paragraph, etc.
- Supports all Wikipedia language editions

Wiktionary, a Multilingual Machine Readable Dictionary

Français
Le dictionnaire libre
856 000+ articles

Tiếng Việt
Từ điển mở
227 000+ từ

Русский
Свободный словарь
137 000+ статей

中文
自由的多语言词典
116 000+ 词条

English
The free dictionary
841 000+ articles

Türkçe
Özgür sözlük
206 000+ madde

Ido
La libera vortaro
137 000+ artikli

Ελληνικά
Το Ελεύθερο Λεξικό
107 000+ λέξεις

Polski
Wolny słownik
93 000+ stron

http://www.wiktionary.org/
Wiktionary – Information Types

- Language
- Etymology
- Pronunciation
- Part of speech
- Derived terms, related terms
- Abbreviations
- Collocations
- Word senses
- Glosses
- Examples
- Synonyms, antonyms, hypernyms, hyponyms
- Translations
- Morphology
- Quotations
- …
JWKTL
Java-based Wiktionary Library

Highlights:

- efficient and structured access to the information encoded in the English, the German, and the Russian Wiktionary language editions
- sense definitions
- part of speech tags
- etymology
- example sentences
- translations
- semantic relations
- … and many other lexical information types.

OmegaWiki, a Multilingual Lexical-Semantic Resource

- A free, multilingual resource
- Over 420,000 expressions in 255 languages
- Over 40,000 language-independent concepts
- Around 3,000 users

Goals
- Overcome Wiktionary’s structural inconsistencies
- Create a resource for translations/synonyms which is easily accessible and maintainable

Consequence: a fixed database schema
- Users can only contribute if they stick to the predefined structure
  - …but the price is a loss in expressiveness

http://www.omegawiki.org/
JOWKL
Java-based OmegaWiki Library

Highlights:

- Fast and efficient access to OmegaWiki
- Direct access to OmegaWiki database dumps, no preprocessing necessary
- Language independent

Available open source: http://code.google.com/p/jowkl/
Part 1: Ingredients and Techniques

Ingredients: Lexical Resources

- Elements of Lexical Resources
- Classic Lexical Resources
- Collaborative Lexical Resources

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing
Lexical Resource Integration – Motivation

Lexical resources are largely different

- Different coverage of words/word senses
- Different information types
  - Encyclopedic vs. linguistic knowledge
  - Syntactic vs. semantic knowledge
- ...

This can significantly influence the performance of an NLP system – Instead of choosing only one (best performing):

Why not combine multiple resources and benefit from all their knowledge?
Linking Lexical Resources at the Word Sense Level

Linking at the word sense level: sense alignment of resources = linking of equivalent senses

Integrated Resource

FrameNet

VerbNet

Wikipedia

Wiktionary [ˈwɪkʃənərɪ] n., a wiki-based Open Content dictionary

GermaNet

IMSlex-Subcat

OmegaWiki

09.2013 | Computer Science Department | UKP Lab | Iryna Gurevych and Judith Eckle-Kohler
Previous Work on Linking Lexical Resources

Linked Lexical Resources:

- Yago, Suchanek et al. (2007)
- SemLink (Palmer, 2009)
- Universal Wordnet (UWN), Gerard de Melo and Gerhard Weikum (2009)
- eXtended WordFrameNet, Laparra and Rigau (2010)
- BabelNet, Navigli and Ponzetto (2010)
- NULEX, McFate and Forbus (2011)
- UBY, Gurevych et al. (2012)
- … many more, e.g., on the Semantic Web
Linking at the Word Sense Level: Example

**to sing**
1. To produce musical or harmonious sounds with one's voice.
2. To express audibly by means of a harmonious vocalization.
3. To confess under interrogation.

**W**
1. Produce tones with the voice
2. divulge confidential information or secrets

**S**
1. To produce harmonious sounds with one's voice.

**Singen**
1. Mit der Stimme harmonische Töne erzeugen.
Sense Alignment of Multiple Resources

Web 2.0
Wikipedia
Wiktionary
OMEGA Wiki
IMSlex-Subcat
FrameNet
VERBNET

09.2013 | Computer Science Department | UKP Lab | Iryna Gurevych and Judith Eckle-Kohler
Case Study: Automatic Word Sense Alignment of Wiktionary and WordNet

Aims:
Create a word sense alignment between Wiktionary and WordNet that comes with
(1) Increased coverage
(2) Enriched sense representations

Meyer, Gurevych (2011)
**Wiktionary vs. WordNet: Differently Developed**

**Wiktionary:** Online lexicon that is collaboratively constructed by a community of Web users

**WordNet:** Semantic network created by psycholinguists at Princeton University (Fellbaum, 1998)

---

**WordNet Search - 3.1**
- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for: plant

Display Options: (Select option to change) [Change](#)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

**Noun**
- S. (n) plant, works, industrial plant (buildings for carrying on industrial labor) "they built a large plant to manufacture automobiles"
- S. (n) plant, flora, plant life (botany) a living organism lacking the power of locomotion
- S. (n) plant (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- S. (n) plant (something planted secretly for discovery by another) "the police used a plant to trick the thieves", "he claimed that the evidence against him was a plant"
Wiktionary vs. WordNet: Different Sense Inventories

**Noun**

- **plant** (plural **plants**)
  1. An **organism** that is not an animal, especially an organism with a defined place in a food chain.
     - The garden had a couple of trees, and a cluster of berries.
  2. (botany) An **organism** of the kingdom Plantae; now specifically, a living organism of the Embryophyta (land plants) or a **b**otanical species of it.
  3. (ecology) Now specifically, a **multicellular** eukaryote that includes chloroplasts in its cells, which have a cell wall.
  4. A **factory** or other industrial or institutional building or facility.
  5. An object placed surreptitiously in order to cause suspicion to fall upon a person.
    - That gun's not mine! It's a plant! I've never seen it before!
  6. Anyone assigned to behave as a member of the **public** during a covert operation (as in a police investigation).
  7. A person, placed amongst an audience, whose role is to cause confusion, laughter etc.
  8. (snooker) A play in which the **cue ball** knocks one (usually red) ball onto another, in order to pot the second; a **set**.

WordNet

- **S:** (n) **plant**, works, industrial plant (buildings for carrying on industrial labor)
  - "they built a large plant to manufacture automobiles"
- **S:** (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)
- **S:** (n) **plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- **S:** (n) **plant** (something planted secretly for discovery by another) "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"
Linking Wiktionary and WordNet at the Sense Level

A two-step approach:
1. Candidate extraction
2. Candidate disambiguation

{plant, works, industrial plant}

WordNet synsets

plant (organism)
plant (factory)
plant (person)

bird (animal)
reddish (color)
to fly (move)

works (factory)
works (machine)

...
Linking Wiktionary and WordNet at the Sense Level

A two-step approach:
1. Candidate extraction
2. Candidate disambiguation

(Niemann & Gurevych, 2011)
Linking Wiktionary and WordNet at the Sense Level

A two-step approach:
1. Candidate extraction
2. Candidate disambiguation

(Niemann & Gurevych, 2011)
Disambiguation: BoW Representation

Synsets are represented by synonyms, gloss, examples.

- Synset
- Hypernyms
- Hyponyms
- Hyper- & hyponyms

- Word
- Gloss
- Usage examples
- Synonyms

Bag-of-words

Wiktionary
['wikjanri] n., a wiki-based Open Content dictionary
Disambiguation: Alignment Classification

- **bag-of-words**
  - semantic relatedness measure
  - **COS**: Cosine similarity
  - **PPR**: Personalized PageRank (Agirre&Soroa, 2009)

- **s < threshold**
  - No alignment!

- **s ≥ threshold**
  - Align this pair of WordNet synset and Wiktionary sense!

---

09.2013 | Computer Science Department | UKP Lab | Iryna Gurevych and Judith Eckle-Kohler 52
Evaluation

Performance of automatic sense alignment framework:
- F1 ~ 0.66, Precision ~ 0.67, Recall ~ 0.65

Open issues to be addressed in future work:
- false negatives “same meaning, but was not aligned”
- Very different wording
- Similar senses but slightly below threshold
- Pointing to another entry rather than a content-based gloss

- false positives “different meaning, but have been aligned”
- Similar wording, but refer to different concepts
- Generic- versus domain-specific vocabulary
## Increased Coverage: Domains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Wiktionary AND WordNet</th>
<th>Only Wiktionary</th>
<th>Only WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>4,465</td>
<td>4,067</td>
<td>12,869</td>
</tr>
<tr>
<td>Chemistry</td>
<td>2,561</td>
<td>8,260</td>
<td>2,268</td>
</tr>
<tr>
<td>Engineering</td>
<td>1,108</td>
<td>940</td>
<td>1,080</td>
</tr>
<tr>
<td>Geology</td>
<td>2,287</td>
<td>2,898</td>
<td>2,479</td>
</tr>
<tr>
<td>Humanities</td>
<td>4,949</td>
<td>2,700</td>
<td>5,060</td>
</tr>
<tr>
<td>IT</td>
<td>439</td>
<td>3,032</td>
<td>557</td>
</tr>
<tr>
<td>Linguistics</td>
<td>1,249</td>
<td>1,011</td>
<td>1,576</td>
</tr>
<tr>
<td>Math</td>
<td>615</td>
<td>2,747</td>
<td>483</td>
</tr>
<tr>
<td>Medicine</td>
<td>3,613</td>
<td>3,728</td>
<td>3,058</td>
</tr>
<tr>
<td>Military</td>
<td>574</td>
<td>426</td>
<td>585</td>
</tr>
<tr>
<td>Physics</td>
<td>1,246</td>
<td>2,835</td>
<td>1,252</td>
</tr>
<tr>
<td>Religion</td>
<td>733</td>
<td>1,154</td>
<td>781</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>3,745</td>
<td>2,907</td>
<td>4,458</td>
</tr>
<tr>
<td>Sport</td>
<td>905</td>
<td>2,821</td>
<td>807</td>
</tr>
</tbody>
</table>
Enriched Sense Representation

Access to complementary information

Synonyms
Gloss
Example sentence
Subsumption hierarchy
Synset organization
...

Wiktionary
[ˈwɪkʃənri] n.,
a wiki-based Open Content dictionary

Pronunciation
Etymology
Usage
Quotations
Related terms
Translations
...

09.2013 | Computer Science Department | UKP Lab | Iryna Gurevych and Judith Eckle-Kohler
The increased coverage and the enriched sense representation yield synergies.

Previously shown:

- Linking FrameNet, VerbNet, and WordNet for semantic parsing (Shi and Mihalcea, 2005)
- Linking VerbNet, FrameNet and PropBank for semantic role labeling (Palmer, 2009)
- Linking WordNet and Wikipedia for word sense disambiguation (Navigli and Ponzetto, 2010)
- Linking WordNet and Wiktionary for measuring verb similarity (Meyer and Gurevych, 2012)

Future work:

- Semantic relatedness, information retrieval, information extraction,…
- Your application?
Part 1: Ingredients and Techniques

Ingredients: Lexical Resources

Elements of Lexical Resources

Classic Lexical Resources

Collaborative Lexical Resources

Integration Technique: Automatic Sense Linking

Interoperability Technique: Standardizing
Interoperability – Motivation

- Different APIs …

JWNL  FN API  JWPL  JWKTL

Wikipedia

Wikipedia

FrameNet

Wiktionary

[ˈwɪkʃənəri] n.,
a wiki-based Open
Content dictionary

09.2013 | Computer Science Department | UKP Lab | Iryna Gurevych and Judith Eckle-Kohler
Integration of Lexical Resources at the Representation Level

Integration at the representation level: standardization of resources

FrameNet
VERBNet
WIKIPEDIA
Integrated Resource
Wiktionary
IMSlex-Subcat
GermaNet
OMEGAWIKI
Heterogeneous Lexical Resources

Different information types

Linguistic terminology

Incompatible data formats
Dimensions of Interoperability

Structural Interoperability:
Uniform lexicon structure – given by the main organizational units
- e.g., harmonizing the synset-based organization (WordNet), the headword-based organization (Wiktionary) and the frame-based organization (FrameNet)

Interoperability at the level of linguistic terminology:
Uniform Data Categories – these are descriptions of the meaning of linguistic terms
- e.g., harmonizing the terms lexical unit (FrameNet) and word sense (Wiktionary)
Standardizing Lexical Resources

ISO Standards for lexical resources:

- ISO 24613:2008 Lexical Markup Framework (LMF)

Standardization of lexical resources according to LMF makes them interoperable:
- at the structural level

- at the level of linguistic terminology
ISO-LMF is an Abstract Standard …

- Classes are given in LMF, but they have no attributes.
- Attributes and their values = linguistic terminology
- In order to make use of ISO-LMF, attributes and their values have to be defined for each class.
- The result is an instantiation of the abstract LMF standard – a lexicon model that can be populated by lexicon data.
An Instantiation of LMF for NLP: UBY-LMF

UBY-LMF

- is an instantiation of LMF that is directly usable
- specifies attributes and their values for all LMF classes used
- extends the LMF standard by two classes: SemanticLabel and Frequency

Many different **types of lexical resources** in UBY-LMF:

- expert built vs. collaboratively constructed
- differently organized:
  - headword-based
  - wordnets
  - lexical resources based on frame semantics
  - subcategorization lexicons
UBY-LMF – Design Principles

Extensibility

- further languages
- further lexicons
- automatically mined information types (e.g. domain labels)

No information loss

- including conflicting information
- e.g. keeping the original sense key in the MonolingualExtRef class
Implementation of UBY-LMF

- Definition of a lexicon model in UML (Unified Modeling Language)

- Implementation of the lexicon model in Java

- Conversion of lexical resources to UBY-LMF according to the lexicon model

- Import of converted resources into an SQL database based on an Object-Relational Mapping using Hibernate

- Export format: XML
Example: SCFs in UBY-LMF – Specification of Syntactic Arguments
### SCFs in UBY-LMF – Example

**GermaNet Syntactic Arguments**

- **syntacticCategory:** nounPhrase
- **grammaticalFunction:** subject
- **syntacticCategory:** prepositionalPhrase
- **grammaticalFunction:** prepositionalComplement
- **Preposition:** -

**VerbNet Syntactic Arguments**

- **syntacticCategory:** nounPhrase
- **grammaticalFunction:** subject
- **syntacticCategory:** prepositionalPhrase
- **grammaticalFunction:** prepositionalComplement
- **Preposition:** about
Homogeneous Lexical Resources

Lexicon model UBY-LMF

ISO-Standard Lexical Markup Framework

Preserves variety of lexical information

Extensible
Outline

Part 1: Ingredients and Techniques

Part 2: Recipes
UBY – the Main Ingredient

Goal: to exploit wide variety of lexical knowledge

Joint work with
Iryna Gurevych, Kostadin Cholakov, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, Tri-Duc Nghiem
Part 2: Recipes

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking
What UBY has to offer

Tools
UBY – Data and Tools

Web Interface

https://uby.ukp.informatik.tu-darmstadt.de/webui/

Database Dumps

http://uby.ukp.informatik.tu-darmstadt.de/uby/

Open Source API (JAVA)

http://code.google.com/p/uby/
UBY Web UI – a Small Glimpse of UBY

- Resources with open licenses
- Selected information types are presented
UBY Web UI – Textual View

Textual View: allows to **list senses across all resources**, to display **sense details** and to perform **sense comparisons**.

OmegaWikien

- **align** (verb)
  - To arrange in a straight line.
- **align** (verb)
  - To bring into cooperation or agreement with a particular group, party, cause, etc.

Wikipedia

- **Align** (noun)
  - Align is a privately held IT solutions company with key competencies in “Technology Transformation Initiatives” – designing, deploying, moving and consolidating technologies and advanced network infrastructures. Specializing in IT Strategy consulting, technology relocation services, network infrastructure design and build-out, contact center technology and IT asset management (ITAM) solutions, the company also provides communications products and integration services for call centers and other facilities requiring large-scale computer telephony installations. Align operates from offices in New York; London; Chicago; Princeton, New Jersey; and Toronto. The company serves customers in such industries as finance, health care and energy. Clients have included Blackrock, BP and Dun and Bradstreet.

WiktionaryEN

- **align** (verb)
  - To form in line; to fall into line.
- **align** (verb)
  - To adjust or form to a line; to range or form in line; to bring into line.

Lexical Information:

- Sense ID: WN_Sense_142794

Semantic Labels:

1. verb.stative

Semantic Information:

- Semantic Predicate
  - Semantic Argument
    - Semantic Role: something

Synset (Original SynsetID:2658867):
- Definition: be or come into adjustment with

Syntactic Information:

1. subject_nounPhrase align complement prepositionalPhrase

Click here for more details...
UBY Web UI – Visual View

Visual view: allows to explore the sense alignments.
The UBY API is **open source** at Google Code:  

Getting Started:

1. Download a UBY database dump  
2. Import the dump into a MySQL database  
3. Start using the UBY API

The UBY API is work in progress!  
Many API methods need to be added – consider contributing!
UBY Database Dumps

Downloads:
http://uby.ukp.informatik.tu-darmstadt.de/uby/

- Resources with **open licenses**
- Database dumps in two different sizes:
  - all UBY resources along with pairwise alignments (without GermaNet, IMSLex), uby_open_<version>
  - all UBY resources along with pairwise alignments without Wikipedia, GermaNet, IMSLex, uby_medium_<version>
UBY API – Getting Started

- Recommended: using the UBY API with Maven
- Currently, it is necessary to use a specific `settings.xml` that configures the access to the public Maven repository maintained by UKP Lab.
- We plan to deploy the UBY API on Maven Central soon!

```xml
<dependency>
  <groupId>de.tudarmstadt.ukp.uby</groupId>
  <artifactId>de.tudarmstadt.ukp.uby.lmf.api-asl</artifactId>
  <version>0.3.0</version>
  <type>jar</type>
  <scope>compile</scope>
</dependency>
```
UBY API – Based on UBY-LMF

UBY-LMF has been implemented in Java. An object-relational mapping by means of the Hibernate framework allows mapping any instance of UBY-LMF to a SQL database.

**UBY is an instance of the UBY-LMF lexicon model:**

- **LexicalResource name:** UBY
  - **Lexicon name:** GermaNet
  - **Lexicon name:** IMSLex
  - **Lexicon name:** WikipediaDE
  - **Lexicon name:** WiktionaryDE
  - **Lexicon name:** OmegaWikiDE
  - **Lexicon name:** WordNet
  - **Lexicon name:** FrameNet
  - **Lexicon name:** WiktionaryEN
  - **Lexicon name:** OmegaWikiEN
  - **Lexicon name:** WikipediaEN
UBY API – Based on UBY-LMF

UBY (as instance of the LexicalResource class) aggregates instances of the Lexicon class.


- Example: querying the names of UBY lexicons

```java
Uby uby = new Uby(dbConfig);
List<String> lexiconNames = uby.getLexiconNames();
```
The uniform representation of resources allows querying across all UBY lexicons.

- Option to filter by POS

- Example: querying lexical entries for a given lemma and POS across all UBY lexicons (here: noun “album”)

```java
Uby uby = new Uby(dbConfig);

List<LexicalEntry> nounEntries = uby.getLexicalEntries("album", EPartOfSpeech.noun, null);
```
UBY API – Uniform Access to UBY Resources

Individual UBY lexicons are queried the same way:

- First, retrieve the lexicon to be queried
- Option to filter by POS

- Example: querying UBY WordNet for the lexical entry of the noun “album”

```java
Uby uby = new Uby(dbConfig);

Lexicon wordNet = uby.getLexiconByName("WordNet");

List<LexicalEntry> lexEntries = uby.getLexicalEntries("album", EPartOfSpeech.noun, wordNet);
```
What UBY has to offer

Planning Guide
Increased Coverage of Lemmas (and their Senses)

UBY lexicons contain different, complementary lemmas:
over 3.08 million unique lemma-POS combinations for English

<table>
<thead>
<tr>
<th>EN Lexicons</th>
<th>noun</th>
<th>verb</th>
<th>adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>699</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>1,630</td>
<td>1,888</td>
<td>430</td>
</tr>
<tr>
<td>3</td>
<td>8,439</td>
<td>1,948</td>
<td>2,271</td>
</tr>
<tr>
<td>2</td>
<td>53,856</td>
<td>4,727</td>
<td>12,290</td>
</tr>
<tr>
<td>1</td>
<td>2,900,652</td>
<td>50,209</td>
<td>41,731</td>
</tr>
<tr>
<td>Σ (unique EN)</td>
<td></td>
<td>3,080,771</td>
<td></td>
</tr>
</tbody>
</table>
Enriched Senses Based on Sense Alignments

- **Definition**
- **SenseExample**
- **SemanticLabel**
- **semanticField**
- **Synset**
- **Subcategorization-Frame**
- **WordNet sense**
- **Wiktionary sense**
- **Definition**
- **SenseExample**
- **SemanticLabel**
- **register**
- **Equivalent**
**Enriched Senses Based on Sense Alignments**

<table>
<thead>
<tr>
<th>Lexicon pair</th>
<th>Languages</th>
<th>SenseAxis</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN–WP-en</td>
<td>EN–EN</td>
<td>50,351</td>
</tr>
<tr>
<td>WN–WKT-en</td>
<td>EN–EN</td>
<td>99,662</td>
</tr>
<tr>
<td>WN–VN</td>
<td>EN–EN</td>
<td>40,716</td>
</tr>
<tr>
<td>FN–VN</td>
<td>EN–EN</td>
<td>17,529</td>
</tr>
<tr>
<td>WP-en–OW-en</td>
<td>EN–EN</td>
<td>3,960</td>
</tr>
<tr>
<td>WP-de–OW-de</td>
<td>DE–DE</td>
<td>1,097</td>
</tr>
<tr>
<td>WN–OW-de</td>
<td>EN–DE</td>
<td>23,024</td>
</tr>
<tr>
<td>WP-en–WP-de</td>
<td>EN–DE</td>
<td>463,311</td>
</tr>
<tr>
<td>OW-en–OW-de</td>
<td>EN–DE</td>
<td>58,785</td>
</tr>
<tr>
<td><strong>UBY</strong></td>
<td><strong>All</strong></td>
<td><strong>758,435</strong></td>
</tr>
</tbody>
</table>

Access to all available, _partly complementary information types_ attached to the aligned senses, e.g., semantic relations, subcategorization frames, encyclopedic or translation information.

For English, ca. **32,000** senses simultaneously take part in _at least two pairwise sense alignments_, i.e. information from 3 UBY lexicons is available.
UBY as Sense Inventory

UBY provides sense inventories at different levels of granularity.

- Example: VerbNet – WordNet alignment
- coarse-grained vs. fine-grained sense inventory

VerbNet:

```
sing
performance-26.7-1-1
```

WordNet:

```
sing
sing%2:36:01
deliver by singing
```
```
sing
sing%2:36:00
produce tones with the voice
```
Main organizational LMF classes
Information Types in UBY – Sense Related

Information types attached to senses

- SenseExample
- SemanticLabel
- Synset
- Subcategorization-Frame
- Semantic-Predicate
- Definition
- SenseRelation
- Equivalent
- Context
Information Types in UBY – Sense Links

Sense alignments between resource pairs

- Sense Axis
  - id
  - senseOne: A
  - senseTwo: B
  - synsetOne
  - synsetTwo
  - senseAxisType

Lexicon X

Sense A

Lexicon Y

Sense B
Wordnets in UBY (Main Information Types)
FrameNet in UBY (Main Information Types)
VerbNet in UBY (Main Information Types)

UBY VerbNet sense != VerbNet sense
Wiktionary and Wikipedia in UBY (Main Information Types)
OmegaWiki in UBY (Main Information Types)
## Mapping of POS for Main Word Classes

<table>
<thead>
<tr>
<th>Resource</th>
<th>Original POS – UBY POS (EPartOfSpeech)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>noun – noun, verb – verb, adjective – adjective</td>
</tr>
<tr>
<td>GermaNet</td>
<td>common noun – nounCommon, proper noun – nounProper, verb – verb, adjective – adjective</td>
</tr>
<tr>
<td>FrameNet</td>
<td>noun – noun, verb – verb, adjective – adjective</td>
</tr>
<tr>
<td>VerbNet</td>
<td>verb – verb</td>
</tr>
<tr>
<td>IMSLex</td>
<td>verb – verbMain, noun – nounCommon, adjective - adjective</td>
</tr>
<tr>
<td>WikipediaEN / DE</td>
<td>common noun – noun, proper noun – noun</td>
</tr>
<tr>
<td>Wiktionary EN / DE</td>
<td>noun – noun, proper noun – nounProper, verb – verb, adjective – adjective</td>
</tr>
<tr>
<td>OmegaWiki</td>
<td>noun – noun, verb – verb, adjective – adjective</td>
</tr>
</tbody>
</table>
Part 2: Recipes

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking
Example Code is Open Source:

http://code.google.com/p/dkpro-tutorials/
Motivation
- The WordNet sense inventory, though standard, has issues, e.g. the overly fine granularity of the senses.

Solution
- UBY as a source for new sense inventories
- richer (coarse-grained) sense inventories
- sense inventories may depend on domains, tasks, etc.

Goal
- Perform sense tagging using UBY

Joint work with Kostadin Cholakov
Occasion

- Sense tagging of text with UBY:
  - to determine which UBY sense occurs in a given text is a **WSD task**

- The sense alignments in UBY help to solve the WSD task, because they enlarge the context (or feature space) for a given UBY word sense
  - e.g., more input for the various types of the Lesk algorithm

Ingredients

- All UBY resources that are aligned

Techniques

- Based on sense linking
Three Ways to Query UBY for WSD

UBY-LMF has been implemented in Java. An object-relational mapping (using Hibernate) allows mapping any instance of UBY-LMF to a SQL database.

The UBY-API `Uby.java` provides methods …

1. to retrieve UBY-LMF objects based on readily available queries
2. that allow you to specify queries yourself
3. Once an UBY-LMF object has been retrieved from the SQL database, the object-relational mapping allows to access all associated objects in the UBY-LMF graph structure.
Retrieving UBY-LMF Objects Based on Readily Available Queries

- **Example:** `getSenseAxisBySense`

```java
for (Sense sense : lexEntry.getSenses()) {
    List<SenseAxis> sas = uby.getSenseAxisBySense(sense);
    for (SenseAxis sa : sas) {
        // DO SOMETHING, SEE CODE EXAMPLES
    }
}
```
Once an UBY-LMF object has been retrieved from the SQL database, the object-relational mapping allows to access all associated objects in the UBY-LMF graph structure.
Accessing Associated Objects in the UBY-LMF Graph Structure

- Example: translations (Equivalent class)

```java
for (Sense sense : lexEntry.getSenses())

    List<Equivalent> eqs = sense.getEquivalents();

    for (Equivalent eq : eqs) {
        System.out.println("- Equivalent: 
+eq.getLanguageIdentifier()
+" : "+eq.getWrittenForm());
    }
```

Accessing Associated Objects in the UBY-LMF Graph Structure

- Example: sense alignments (SenseAxis class)

```java
for (SenseAxis sa : sas) {

    if (sa.getSenseOne().getId().matches("WktEN.*")) {
        Sense wktSense = sa.getSenseOne();
        String wktVerb =
            wktSense.getLexicalEntry().getLemmaForm();
    }
    else if (sa.getSenseTwo().getId().matches("WktEN.*")) {
        Sense wktSense = sa.getSenseTwo();
        String wktVerb =
            wktSense.getLexicalEntry().getLemmaForm();
    }
}
```
Using the UBY API to Specify Queries Yourself

- Criteria API provided by Hibernate

- Example: retrieving the mapping of syntactic and semantic arguments from VerbNet

```java
Uby uby = new Uby(dbConfig);
session = uby.getSession();

Criteria criteriaSynSem =
    session.createCriteria(SynSemArgMap.class);
List<SynSemArgMap> SynSemArgMaps =
    criteriaSynSem.list();
for (SynSemArgMap synSem : SynSemArgMaps) {
    SynargSemargMap.put(
        synSem.getSyntacticArgument(),
        synSem.getSemanticArgument());
}
```
Good to Know – Extraction of Bulk Data from UBY

To extract all instances of LexicalEntry, Sense, SenseAxis, always use the Iterator methods.

- **Example:** `lexicalEntryIterator`
  - Option to filter by POS and Lexicon

```java
Iterator<LexicalEntry> lexicalEntryIterator = uby.getLexicalEntryIterator(null, lex);

while (lexicalEntryIterator.hasNext()) {
    LexicalEntry le = lexicalEntryIterator.next();
    for (Sense s: le.getSenses()) {
        System.out.println(s.getLexicalEntry().getLemmaForm());
    }
}
```
Part 2: Recipes

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking
UBY as UIMA Resource

Motivation

- Using UBY (or a UBY lexicon) as interchangeable resource in UIMA-based NLP pipelines

Solution

- Specifying UBY as UIMA resource based on uimaFIT

Goal

- Retrieving information from UBY in UIMA annotators

Joint work with Richard Eckart de Castilho
UBY as UIMA Resource – Recipe

Occasion

- UIMA annotator needs to retrieve information from UBY

Ingredients

- UBY
- uimaFIT

Techniques

- Based on standardization
Major goal: transform **unstructured** information to **structured** information

... in order to discover knowledge that is relevant to an end user

- **Component-based** architecture for analysis of unstructured content like text, video, audio
- Originally developed at IBM – today an Apache project
- Used in commercial as well as educational contexts
  - LanguageWare, Watson (IBM)
  - **uimaFIT** (University of Colorado, now: Apache UIMA uimaFIT™)
  - **DKPro Core** (TU Darmstadt)
  - many more...

How it works: think of UIMA components as machines in an assembly line
Analysis Levels in Text Processing

- Unstructured
- Segmentation
- Morphology
- Syntax
- Semantics

Structured
UIMA Example Pipeline

Segmentation
Morphology
Syntax
Semantics

Collection Reader
Segmenter
POS Tagger
Lemmatizer
UbyAnnotator
CAS Consumer
Interoperable Lexical Resources can be used as UIMA resources that may be shared between UIMA components.
uimaFIT

uimaFIT is an „add-on“ for UIMA, it simplifies typical development tasks and allows for rapid and easy development of NLP processing pipelines.


For instance, uimaFIT provides annotations for component configuration:

- @ConfigurationParameter annotation
- @ExternalResource annotation for (shared) resource management

To specify UBY as UIMA resource based on uimaFIT, we use

- the base class for external resources Resource_ImplBase
- the interface ExternalResourceLocator
- the @ExternalResource annotation
@ExternalResource Annotation

uimaFIT provides an annotation-based mechanism for the **configuration of UIMA resources** in annotator components.

Example: @ExternalResource annotation used in a UbyAnnotator component

```java
public static final String PARAM_UBY_RESOURCE = "uby";
@ExternalResource(key = PARAM_UBYRESOURCE)
Uby uby;
```
UBY Annotator – Example Use Cases

A UBY annotator can add information from UBY to the CAS in order to enrich the context of a token.

- What can be annotated at the token level?
  - Semantic label information, e.g., semantic field, domain, register
  - Lexical information attached to related senses, e.g., synonyms, hypernyms
  - Translations

- Example use cases include
  - processing of text corpora (document collections) in order to perform knowledge-based sense tagging, e.g., using Lesk-based WSD
  - processing of example sentences from lexical resources in order to acquire more fine-grained lexical information, e.g., subcategorization frames
UBY as Resource – Pipeline Configuration

```java
AnalysisEngineDescription analysisEngine =
    createEngineDescription(
        UbyAnnotator.class,
        UbyAnnotator.PARAM_UBY_RESOURCE,
        createExternalResourceDescription(
            UbyResourceLocator.class,
            UbyResourceLocator.PARAM_URL, "localhost/uby_open",
            UbyResourceLocator.PARAM_DRIVER, "com.mysql.jdbc.Driver",
            UbyResourceLocator.PARAM_DRIVER_NAME, "mysql",
            UbyResourceLocator.PARAM_USERNAME, "user",
            UbyResourceLocator.PARAM_PASSWORD, "pass"
        )
    );
```
Part 2: Recipes

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking
Semantic Tagging with UBY

Motivation

- WordNet provides semantic field information that is often used as coarse-grained semantic feature in NLP
- However, the coverage of WordNet is limited
  - This is especially true for the vocabulary (i.e., lemmas) covered: **WordNet lacks vocabulary** for particular domains, e.g., IT domain

Solution

- Exploiting the **increased lemma coverage** offered by UBY

Goal

- Broad-coverage semantic tagging of text with semantic field information

Joint work with Richard Eckart de Castilho
Semantic Tagging with UBY – Recipe

Occasion
- Broad-coverage semantic tagging of text with semantic field information

Ingredients
- UBY version of WordNet / GermaNet
- UBY
- DKPro Core and uimaFIT

Techniques
- Based on standardization
Usage of Semantic Field Information in NLP

WordNet provides coarse-grained **semantic field information** through its lexicographer file names.

WordNet semantic field information has widely been used in many NLP tasks and applications, including Information Extraction and text classification.

- Semantic field information from WordNet has also been called *supersenses* and has been applied for *supersense tagging*, e.g.
- (Ciaramita and Johnson, 2003) and (Qiu et al., 2011) each present a framework for classifying words not in WordNet into the WordNet semantic fields
- (Ciaramita and Altun, 2006) describe a supersense tagger based on sequence labeling
UIMA Example Pipeline for Semantic Tagging

Segmentation
Morphology
Syntax
Semantics

Collection Reader
Segmenter
POS Tagger
Lemmatizer
UbySemanticFieldAnnotator
CAS Consumer
DKPro Core – Linguistic Annotation of Text


DKPro Core builds on uimaFIT.

**Integrated Tools**
- TreeTagger
- OpenNLP
- Stanford NLP
- JWordSplitter
- Language Tool
- MaltParser
- ...

**Supported Formats**
- Text
- PDF
- TEI XML, BNC XML
- SQL Databases
- Google web1t n-grams
- ...

Interface SemanticTagProvider

- Interface to create various UIMA resources that provide semantic tags:
  SemanticTagProvider

- The interface SemanticTagProvider can be used to retrieve semantic tag information from:
  - simple key value files
  - UBY
  - many other (linked) lexical resources, such as BabelNet, UWN, ...

SemanticTagProvider offers the method getSemanticTag:

String getSemanticTag(Token token)
UIMA Resource UbySemanticFieldResource

UbySemanticFieldResource

- **extends** Resource_ImplBase **and implements** SemanticTagProvider

- follows the first sense (WordNet) / random sense (other UBY resources) heuristic

How to use UbySemanticFieldResource in annotator components:

```java
public static final String PARAM_UBY_SEMANTIC_FIELD_RESOURCE = "ubySemanticFieldResource";

@ExternalResource(key = PARAM_UBY_SEMANTIC_FIELD_RESOURCE)
private UbySemanticFieldResource ubySemanticFieldResource;
```
How to use UbySemanticFieldAnnotator in a pipeline:

```java
AnalysisEngineDescription processor = createEngineDescription(
    createEngineDescription(
        UbySemanticFieldAnnotator.class,
        UbySemanticFieldAnnotator.PARAM_UBY_SEMANTIC_FIELD_RESOURCE,
        createExternalResourceDescription(
            UbySemanticFieldResource.class,
            UbySemanticFieldResource.PARAM_URL, "localhost/uby_open",
            UbySemanticFieldResource.PARAM_DRIVER_NAME, "mysql",
            UbySemanticFieldResource.PARAM_USERNAME, "root",
            UbySemanticFieldResource.PARAM_PASSWORD, "pass"
        )
    )
);```

Part 2: Recipes

Tools and Planning Guide: What UBY has to offer

Mixed Starters: How to query UBY (for WSD)

Appetizer: UBY as UIMA resource

Main Dish: UBY for UIMA-based Semantic Tagging

Dessert: Cross-lingual verb sense linking
Cross-lingual Verb Sense Linking

Motivation

- German resources that are publicly available for research lack semantic role and selectional preference information for verbs.

Solution

- Exploiting the standardized format for subcategorization frames (SCFs) in English (EN) and German (DE) provided by UBY-LMF in order to perform a cross-lingual linking of verb senses and their SCFs

Goal

- Enriching verb senses in the German resources GermaNet and IMSLex
Cross-lingual Verb Sense Linking – Recipe

Occasion

- Enriching verb senses in GermaNet and IMSLex with semantic role and selectional preference information
- Using this information, e.g., in WSD

Ingredients

- UBY versions of IMSLex, GermaNet, WordNet, VerbNet

Techniques

- Based on sense linking and standardization
Standardized Format for SCFs in English (EN) and German (DE)

**Requirement:** flexible format for SCFs in EN and DE, applicable both across languages and at the language-specific level

- **Across languages:** in order to automatically detect correspondences
- **Language-specific:** preserve distinctions between fine-grained EN and DE SCFs in order to automatically identify them as well

**UBY-LMF** – based on ISO-Standards – meets this requirement

- ISO 24613:2008 Lexical Markup Framework (LMF)
SCFs in UBY-LMF – Specification of Syntactic Arguments
SCFs in UBY-LMF – Example

GermaNet Syntactic Arguments

- **syntacticCategory**: nounPhrase
- **grammaticalFunction**: subject
- **syntacticCategory**: prepositionalPhrase
- **grammaticalFunction**: prepositionalComplement
- **Preposition**: -

VerbNet Syntactic Arguments

- **syntacticCategory**: nounPhrase
- **grammaticalFunction**: subject
- **syntacticCategory**: prepositionalPhrase
- **grammaticalFunction**: prepositionalComplement
- **Preposition**: about

---

**lachen**

**laugh**

NN.Pp

NP V PP {about}
## Syntactic Arguments in UBY-LMF: Attributes

- **syntacticCategory**
- **grammaticalFunction**
- **case**
- **determiner**
- **number**
- **verbForm**
- **tense**
- **complementizer**
- **preposition**
- **lexeme**

Defined for all syntactic arguments.

Language-specific attribute values.
Cross-lingual Linking of Verb Senses

IMSLex-Subcat (Eckle-Kohler, 1999)

WordNet Verb lemmas

Interlingual Index (ILI)

GermaNet

UBY
IMSLex-Subcat

Verbs from IMSLex-Subcat which can be used with zu-infinitive or complement clause: **784 verbs**

Interlingual Index (ILI)

UBY

VerbsNet Verb lemmas

IMSLex-Subcat

WordNet Verb lemmas

GermaNet
How to Exploit UBY-LMF Compliant SCFs: Linking of IMSLex-Subcat and GermaNet

- Same number of syntactic arguments
- syntacticCategory, case, and complementizer values are compatible

**lexicon entries are linked**

44 (out of 784) IMSLex verbs do not occur in GermaNet
How to Exploit UBY-LMF Compliant SCFs: Linking of IMSLex-Subcat und VerbNet

- Same number of syntactic arguments
- syntacticCategory values are compatible
- Based on translation via ILI

(140 WordNet verbs do not occur in VerbNet)

<table>
<thead>
<tr>
<th>syntacticCategory</th>
<th>IMSLex-Subcat</th>
<th>lemma</th>
<th>SCF</th>
<th>lemma EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>gramm.Function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>case</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>determiner</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verbForm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tense</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>complementizer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>preposition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lexeme</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

72.68% Precision

Semantic roles, Selectional preferences ...
Future Work: Cross-lingual Verb Sense Linking on the Semantic Web

lemonUby is the Semantic Web version of UBY
http://lemon-model.net/lexica/uby/

- lemon: lexicon model for the Semantic Web
- mapping from UBY-LMF to lemon

lemonUby

- comprises a subset of UBY resources
- lemonUby is linked with other language resources in the Linguistic Linked Open Data (LLOD) cloud:
  - lexical resources: WordNet 3.0, WordNet 2.0, Wiktionary
  - linguistic terms in lemonUby are linked to the Ontologies of Linguistic Annotations (OLiA) (Chiarcos, 2012)
Take Home Messages

Automatically linking lexical resources at the sense level
- is a difficult NLP task that involves WSD
- benefits from lexical resource standardization

Many linked lexical resources exist and their potential for NLP is waiting to be tapped.

Unique features of UBY are
- the wide variety of lexical resources it integrates
- its ISO-compliant standardized format
Thank You!

Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt
Selected References –
Classic Lexical Resources

WordNet, GermaNet

FrameNet

VerbNet

IMSLex-Subcat
Selected References – Collaborative Resources

Wikipedia, Wiktionary, OmegaWiki


Selected References –
Linked Lexical Resources I


Selected References –
Linked Lexical Resources II

Selected References –
Linking Lexical Resources I


Selected References – Linking Lexical Resources II


Selected References – Standardizing Lexical Resources I


Selected References –
Standardizing Lexical Resources II

Selected References – Semantic Tagging