Wordnet - a multilingual semantic lexicon

Darja Fišer
1. What is wordnet & Why it’s good for
2. 3 approaches to wordnet development
3. Automatic extension of wordnet
4. Automatic cleaning of noisy synsets
5. Browsing, editing and visualization of wordnet
6. Conclusions & future plans
1. Background & Motivation
(What is wordnet & Why it’s good for)
What are semantic lexicons

- computer databases of human knowledge about our words & worlds
- explicit structural, semantic & relational information
- vocabulary is organized according to the meaning (tree > birch, car ~ automobile)
- more structured than dictionaries but less formal than ontologies
Wordnet

Noun

- **S:** (n) **bug#1** (general term for any insect or similar creeping or crawling invertebrate)
  - *direct hypernym / inherited hypernym / sister term*
    - **S:** (n) **insect#1** (small air-breathing arthropod)
  - *derivationally related form*
- **S:** (n) **bug#2**, **glitch#1** (a fault or defect in a computer program, system, or machine)
- **S:** (n) **bug#3** (a small hidden microphone; for listening secretly)
- **S:** (n) **hemipterous insect#1**, **bug#4**, **hemipteron#1**, **hemipteran#1** (insects with sucking mouthparts and forewings thickened and leathery at the base; usually show incomplete metamorphosis)
- **S:** (n) **microbe#1**, **bug#5**, **germ#3** (a minute life form (especially a disease-causing bacterium); the term is not in technical use)

Verb

- **S:** (v) **tease#1**, **badger#1**, **pester#1**, **bug#1**, **beleaguer#1** (annoy persistently)
  "The children teased the boy because of his stammer"
- **S:** (v) **wiretap#1**, **tap#5**, **intercept#2**, **bug#2** (tap a telephone or telegraph wire to get information) "The FBI was tapping the phone line of the suspected spy": "Is this hotel room bugged?"
**Wordnet**

**Noun**
- **S**: (n) bug#1 (general term for any insect or similar creeping or crawling invertebrate)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - **S**: (n) insect#1 (small air-breathing arthropod)
  - *derivationally related form*
- **S**: (n) bug#2, glitch#1 (a fault or defect in a computer program, system, or machine)
- **S**: (n) bug#3 (a small hidden person; for listening secretly)
- **S**: (n) hemipterous insect#1, bug#4, hemipteron#1, hemipteron#1 (insects with sucking mouthparts and forewings thickened and leathery at the base; usually show incomplete metamorphosis)
- **S**: (n) microbe#1, bug#5, germ#3 (a minute life form (especially a disease-causing bacterium); the term is not in technical use)

**Verb**
- **S**: (v) tease#1, badger#1, pester#1, bug#1, beleaguer#1 (annoy persistently) "The children teased the boy because of his stammer"
- **S**: (v) wiretap#1, tap#5, intercept#2, bug#2 (tap a telephone or telegraph wire to get information) "The FBI was tapping the phone line of the suspected spy"; "Is this hotel room bugged?"
1. milk -- (Gastronomy) a white nutritious liquid secreted by mammals and used as food by human beings
   ➞ pasteurized_milk -- (Gastronomy) milk that has been exposed briefly to high temperatures to destroy
   ➞ cows'_milk -- (Gastronomy) milk obtained from dairy cows
   ➞ yak's_milk -- (Gastronomy) the milk of a yak
   ➞ goats'_milk -- (Gastronomy) the milk of a goat
   ➞ acidophilus_milk -- (Gastronomy) milk fermented by bacteria; used to treat gastrointestinal disorders
   ➞ pasturized_milk -- (Gastronomy) subjected
   ➞ raw_milk -- (Gastronomy) unpasteurized
   ➞ scalded_milk -- (Gastronomy) milk heated
   ➞ homogenized_milk -- (Gastronomy) milk
   ➞ certified_milk -- (Gastronomy) from dairy
   ➞ powdered_milk, dry_milk, dried_milk,
   ➞ evaporated_milk -- (Gastronomy) milk
   ➞ condensed_milk -- (Gastronomy) sweet
   ➞ skim_milk, skimmed_milk -- (Gastronomy)
   ➞ whole_milk -- (Gastronomy) milk from
   ➞ buttermilk -- (Gastronomy) residue from
   ➞ chocolate_milk -- (Gastronomy) milk fl

1. leche -- (Gastronomy) [a white nutritious liquid secreted by mammals and used as food by human beings
   ➞ leche_pasteurizada -- (Gastronomy) [milk that has been exposed briefly to high
   ➞ leche_de_vaca -- (Gastronomy) [milk obtained from dairy cows]
   ➞ leche_de_yac -- (Gastronomy) [the milk of a yak]
   ➞ leche_de_cabra -- (Gastronomy) [the milk of a goat]
   ➞ acidophilus_milk] -- (Gastronomy) [milk fermented by bacteria; used to treat gastrointestinal disorders
   ➞ leche_pasteurizada -- (Gastronomy) subjected to carefully controlled heating
   ➞ leche_cruda -- (Gastronomy) [unpasteurized milk]
   ➞ leche_hervida -- (Gastronomy) [milk heated almost to boiling]
   ➞ leche_homogeneizada -- (Gastronomy) [milk with the fat particles broken up
   ➞ leche_certificada -- (Gastronomy) [from dairies regulated by an authorized m
   ➞ leche_en_polvo -- (Gastronomy) [dehydrated milk]
   ➞ leche_evaporada -- (Gastronomy) [milk concentrated by evaporation]
   ➞ leche_condensada -- (Gastronomy) [sweetened evaporated milk]
   ➞ leche_descremada, leche_desnatada -- (Gastronomy) [milk from which the cr
   ➞ whole_milk] -- (Gastronomy) [milk from which no constituent (such as fat) h
   ➞ low-fat_milk] -- (Gastronomy) [milk from which some of the cream has been
   ➞ suero_de_la_leche -- (Gastronomy) [residue from making butter from sour ra
   ➞ leche_con_chocolate -- (Gastronomy) [milk flavored with chocolate syrup]
Why do we need them?

• bridge between language & knowledge
  ‣ semantic normalization (bug, pester)
  ‣ disambiguation (bug_insect, bug_defect)

• HLT applications:
  ‣ search engines
  ‣ machine translation
  ‣ document classification
  ‣ information extraction
  ‣ text summarisaton
Why automatic construction?

• needs:
  ‣ 1 lexical entry ~30 min
  ‣ lexicon size ~100,000 entries
  ‣ ~50,000 hours / ~2,000 days / ~6 years

• aims:
  ‣ speed up
  ‣ simplify
  ‣ lower costs
  ‣ recycle
2. Wordnet development
   (3 approaches)
Research goals

• develop methodology & test 3 different multilingual approaches

• expand vs. merge approach
  ▸ translational relation
  ▸ parallel wordnets
1. Dictionary approach

- Rigau idr. (1998)
  - bilingual dictionary
  - obvious choice
  - rich vocabulary
  - ready-made translations
  - different sense inventory
  - mapping to wn synsets not trivial
## 1. Dictionary approach

### Noun
- **S:** (n) bug#1 (general term for any insect or similar creeping or walking invertebrate)
  - direct hypernym / inherited hypernym / sister term
  - **S:** (n) insect#1 (small air-breathing arthropod)
  - derivationally related form
- **S:** (n) bug#2, glitch#1 (a fault or defect in a computer program or machine)
- **S:** (n) bug#3 (a small hidden microphone; for listening secretly)
- **S:** (n) hemipterous insect#1, bug#4, hemipteran#1, hemiptera#2 (insects with sucking mouthparts and forewings thickened and leathery; bugs usually show incomplete metamorphosis)
- **S:** (n) microbe#1, bug#5, germ#3 (a minute life form (especially causing bacterium); the term is not in technical use)

### Verb
- **S:** (v) tease#1, badger#1, pester#1, bug#1, beleaquer#1 (annoy)
  - "The children teased the boy because of his stammer" 
- **S:** (v) wiretap#1, tap#5, intercept#2, bug#2 (tap a telephone or electrical wire to get information) "The FBI was tapping the phone line of the suspected spy"; "Is this hotel room bugged?"
1. Dictionary approach

**wordnet A** → **dictionary A-B** → **wordnet B**

**Serbian WN**

- konac, kraj, svršetak, završetak

**Srp-Slo dict.**

- konac: izid, iztek, konec, končanje, kraj, krajnik, obrobje, nit, sklep, sukanec, zaključek, zatrep

**Slovene WN**

- izid, iztek, konec, končanje, kraj, sklep,
2. Corpus approach

- Diab (2004)
  - multilingual parallel corpora
  - existing wordnets for several languages
  - automatic sense-assignment to polysemous words
  - more pre-processing
  - limited to single-word literals
2. Corpus approach

- sentence alignment
2. Corpus approach

- POS-tagging
- lemmatization
- word alignment
- lexicon extraction
## 2. Corpus approach

<table>
<thead>
<tr>
<th>EN</th>
<th>CS</th>
<th>RO</th>
<th>BG</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>word</td>
<td>word</td>
<td>word</td>
<td>word</td>
</tr>
<tr>
<td>party</td>
<td>strana</td>
<td>partid</td>
<td>партия</td>
<td>stranka</td>
</tr>
<tr>
<td>party</td>
<td>večírek</td>
<td>petrecere</td>
<td>забава</td>
<td>zabava</td>
</tr>
<tr>
<td>army</td>
<td>armáda</td>
<td>armată</td>
<td>армия</td>
<td>armada</td>
</tr>
<tr>
<td>army</td>
<td>armáda</td>
<td>armată</td>
<td>армия</td>
<td>vojska</td>
</tr>
</tbody>
</table>
## 2. Corpus approach

<table>
<thead>
<tr>
<th>EN</th>
<th>CS</th>
<th>RO</th>
<th>BG</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>wn</td>
<td>word</td>
<td>wn</td>
<td>word</td>
</tr>
<tr>
<td>party</td>
<td>01</td>
<td>strana</td>
<td>01</td>
<td>partid</td>
</tr>
<tr>
<td>party</td>
<td>02</td>
<td>večírek</td>
<td>02</td>
<td>petrecere</td>
</tr>
<tr>
<td>army</td>
<td>03</td>
<td>armáda</td>
<td>03</td>
<td>armată</td>
</tr>
<tr>
<td>army</td>
<td>03</td>
<td>armáda</td>
<td>03</td>
<td>armată</td>
</tr>
</tbody>
</table>
## 2. Corpus approach

<table>
<thead>
<tr>
<th>EN</th>
<th>CS</th>
<th>RO</th>
<th>BG</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>wn</td>
<td>word</td>
<td>wn</td>
<td>word</td>
</tr>
<tr>
<td>party</td>
<td>01</td>
<td>strana</td>
<td>01</td>
<td>partid</td>
</tr>
<tr>
<td>11</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>party</td>
<td>02</td>
<td>večírek</td>
<td>02</td>
<td>petrecere</td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>army</td>
<td>03</td>
<td>armáda</td>
<td>03</td>
<td>armată</td>
</tr>
<tr>
<td>16</td>
<td>99</td>
<td>55</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>army</td>
<td>03</td>
<td>armáda</td>
<td>03</td>
<td>armată</td>
</tr>
<tr>
<td>33</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Encyclopedic approach

- Navigli & Ponzetto (2010)
  - extensive & multilingual publicly available resource
  - multi-word terms
  - domain-specific terms
  - small size of Slovene Wikipedia (64,000 articles << 2.5 million articles in English)
  - fake monosemy
3. Encyclopedic approach

**Crop rotation**
From Wikipedia, the free encyclopedia

"Fallow" redirects here. For other uses, see Fallow (disambiguation).

Crop rotation or Crop sequencing is the practice of growing a series of crops on the same piece of land in an organized manner...

**Kolobarjenje**
Iz Wikipedije, proste enciklopedije

Kolobarjenje (tudi kolobar) je metoda, pri kateri se vrtnine različnih talnih škodljivcev, ki napadajo točno določene vrste bolezni.
Results

- no. of synsets: 16.886
- no. of literals: 19.582
- % of PWN: 15%
- % of BCS1 & BCS2: 100%
- % of nouns: 91%
- % of MWE: 43%
- 1 literal/synset: 66%
- synset length: 1,16
- longest synset: 16 literals (goljufati)
Analysis

• domains:
  ‣ factotum 25 % (dictionary & corpus)
  ‣ zoology 17 % (wiki)
  ‣ botany 13 % (wiki)
  ‣ biology 7 % (wiki)
  ‣ (agriculture ~330 synsets)

• semantic relations:
  ‣ hypernymy 46 %, 91 % for nouns
  ‣ complete chains 46 %
  ‣ longest chain 16 nodes (telica)
Vocabulary coverage

Noun senses

- 59% not in sloWNet
- 23% monosemous
- 18% polysemous

Noun frequency

- <3: 28% not in sloWNet, 72% not in sloWNet
- <30: 65% not in sloWNet, 35% not in sloWNet
- >30: 91% not in sloWNet, 9% not in sloWNet
3. Wordnet extension
(with Benoît Sagot, INRIA, France)
Motivation & approach

• current wordnet reliable but small
• available resources not used to their full potential
• idea:
  ‣ extract all possible translation equivalents
  ‣ use current wordnet as “copper standard”
  ‣ train a Maximum Entropy classifier
  ‣ add (synset,literal) pairs that are above threshold
• ~ 300,000 translation equivalents extracted
• features used for the classifier:
  ‣ semantic proximity (wn vs. corpus - SemanticVectors)
  ‣ number of sources yielding the same (synset,literal) pair
  ‣ level of polysemy
  ‣ number of tokens in literal
• threshold was set empirically to 0.1
Evaluation of the results

- ~ 68,000 (literal, synset) pairs were added to wn
- 63,010 (63%) new (literal, synset) pairs were added
- 25,102 synsets that were empty before now have at least 1 literal
- sloWNet 3.0: 82,721 (literal,synset) pairs & 42,919 synsets
- manual evaluation of 400 (literal, synset) pairs above the threshold: 64% accuracy
- automatic evaluation against goldstandard: 85% accuracy
3. Cleaning noisy synsets
(work in progress)
Motivation & approach

• current wordnet large but noisy
• biggest errors are due to poor wsd (organ_body, organ_instrument)
• idea: rank a (noisy) list of synonym candidates with distributional methods for detecting semantic similarity between words
• hypothesis: lexemes tend to co-occur in corpora with other semantically related lexemes, as made explicit by relations between synsets in a wordnet
4. sloWTool
(browsing, editing & visualization)
Motivation

• available tools:
  ‣ Princeton WordNet: Princeton WordNet Browser
  ‣ EuroWordNet: Polaris & Periscope
  ‣ BalkaNet: DEBVisDic

• ideal tool:
  ‣ freely available, platform-independent, on-line
  ‣ all-in-one tool (browsing, editing & visualization)
  ‣ support for standardized formats (e.g. LMF)
  ‣ support for multilingual scenarios
  ‣ easy integration of third-party resources (e.g. domains, coarse-grained sense clusters, images)
Browsing

sloWTool

- Search field
- No. of results
- Main menu
- Copy URL
- Select language
Browsing

- **Synonyms**
  - English: *dactyl, digit*
  - Slovenian: *daktil, prst*

- **Definition**
  - English: A finger or toe in human beings or corresponding body part in other vertebrates.

- **Semantic Relations**
  - Derived: *digitalen, digital*
  - Holonym: *vretenčar, craniate, vertebrate*
  - Hypernym: *ud, del, član, ud, okončina, končina, ekstremiteta, member, appendage, extremity*

- **Example**
  - English: An external body part that projects from the body.
  - Usage: It is important to keep the extremities warm.

- **Meta Info**
  - POS: Noun
  - ID: eng-30-05566097-n
  - Domain: Anatomy

- **Edit Info**
  - Stamp: Darja 2008-01-01 00:00:00
Editing

Register & login

Name: 
Surname: 
UserName: 
Email: 
Password: 
Retype password: 

POS: Noun ID: eng-30-05297523-n BCS: 1 DOMAIN: anatomy

SYNONYM (SLV): organ, orgle

SYNONYM (ENG): organ

DEFINITION: a fully differentiated structural and functional unit in an animal that is specialized for some particular function

→ [ENG_DERIVATIVE]: organski, organic
→ [HYPERNYM]: del telesa, body part

STAMP: darja 2008-01-01 00:00:00
Visualization

tla, zemlja, tla, ozemljitev
prst, zemlja
črnica, humus,...

prst, zemlja

prst, zemlja

prst

dolžinska enota

prst

measure

rokovica

prst

glove

pokrivalo

prst

fingert

dlan, taca, šana
ekstremiteta, končina,...

ekstremiteta, končina,...

prst, dakt

otipati
5. Conclusions
(and more future work)
nl.ijs.si/slownet

sloWNet
Slovene Wordnet

What is sloWNet?
sloWNet is a lexico-semantic resource for Slovene, in which words that have the same meaning (literals) are organized into sets of synonyms (synsets). Synsets are linked into a semantic network with various lexical and semantic relations.

The wordnet family:
The first wordnet was developed for English in the 1980's at Princeton University and it became one of the most popular resources for tasks in the field of automatic understanding of natural language. Wordnets for other languages soon followed in projects, such as EuroWordNet, BalkaNet and MultiWordNet. Wordnets for 50 different languages are currently registered with the Global WordNet Association.

How was sloWNet built?
sloWNet was built automatically. The creation process consisted of three stages:
1. Core wordnet
   A bilingual dictionary was used to translate basic concepts into Slovene. The translations were then checked and corrected by hand.
2. Polysemous words
   Polysemous words were dealt with an approach in which a parallel corpus for five languages was word-aligned and the extracted multilingual lexicon was disambiguated with the existing wordnets for these languages.
3. Monosemous words
   Equivalents for monosemous words were found in open-source resources, such as Wikipedia and Eurovoc thesaurus.

What is in sloWNet?
Number of entries
sloWNet currently contains about 20,000 unique literals which are organized into almost 17,000 synsets.

Sources of entries

• XML format
• CC licence
• viewing & editing in DEBVisDic
Conclusions

• advantages of the model:
  ‣ faster & easier construction
  ‣ modularity
  ‣ language-independent (WOLF, Fišer & Sagot 2008)

• disadvantages of the model
  ‣ fine-grained senses
  ‣ inherited inconsistencies
  ‣ English-centered
Future plans

• further development of sloWNet:
  ‣ cleaning & refinement
  ‣ add domain-specific terminology
  ‣ verify the sense inventory in a monolingual reference corpus

• use of sloWNet:
  ‣ semantic annotation of a corpus (on-going)
  ‣ automatic word-sense disambiguation
  ‣ use sloWNet in NLP tasks & applications
Thank you!