Yet Another RussNet: Spinning-in-Progress

P. Braslavski, D. Ustalov, M. Mukhin, Y. Kiselev
Ural Federal University
Yekaterinburg, Russia
Outline

• Introduction
• Related Work
• Structure
• Implementation
• Evaluation
• Conclusion
Introduction

• A thesaurus is a critical resource for successful NLP and AI applications.
  • No open source thesaurus for Russian.
• Yet Another RussNet, started in 2013, is aimed at creation of such one.
  • Crowdsourcing is used.

http://russianword.net/en/
Related Work: Thesauri

- **Created by experts:**
  - RussNet, RuThes(-lite), UNL.

- **Crowdsourcing:**
  - Russian Wiktionary.

- **Automatically derived:**

- **Other Slavic languages:**
  - plWordNet, BulNet, etc.
Related Word: Crowdsourcing

• Three genres of crowdsourcing.
• Games with a purpose.
  • Ex.: Phrases Detectives, JeuxDeMots, etc.
• Mechanized labor.
  • Ex.: Mechanical Turk, CrowdFlower, etc.
• Wisdom of the crowds.
  • Ex.: Wikipedia, Wiktionary, etc.
YARN Structure

• **YARN** is conceptually similar to **PWN**.
• Words have grammatical features.
• Synsets may contain glosses.
• Words in synsets can have definitions, usage examples, and labels.
• Each synset may belong to a domain.
YARN Structure: Relations

• Synsets are linked to each other primarily via *is-a* relations.

• Other relations:
  • meronymy *(part-of)* between synsets,
  • antonymy between lexical senses.

• We elaborated 4-5 top levels for each part of speech.
YARN Structure: Raw Data

• YARN is not created from scratch.

• The following “raw data” are used:
  • Wiktionary (as the core),
  • Wikipedia redirects,
  • UNLDC,
  • Russian National Corpus statistics.

• The goal of YARN is to refine these data to make a successful resource.
Current State of YARN

• More than 200 people have taken part in the synset assembly.
• The resource comprises 100K+ words and 46K+ synsets under CC BY-SA.
• WotC: https://russianword.net/editor.
Synsets & Users

- The graph on the right shows the distribution of users across different number of edits:
  - (0, 10] range: 0 users
  - (10, 100] range: 25 users
  - (100, 500] range: 75 users
  - (500, 1K] range: 50 users
  - (1K, +Inf] range: 10 users

- The graph on the left depicts the distribution of synsets based on the number of words:
  - 1 word: 0 synsets
  - 2 words: 5 synsets
  - 3 words: 10 synsets
  - 4 words: 20 synsets
  - 5+ words: 200 synsets

- The number of users decreases as the number of edits increases, with the highest concentration in the (100, 500] range.
Implementation Details

• Web-based app: **Ruby on Rails**.
• Data are internally stored in a **PostgreSQL** database.
  • Aggregated action chunks are tracked.
• Export formats: XML, Turtle, CSV.
• Import formats: XML, CSV.
• Schema is (somewhat) similar to LMF.
User Interface

“bedside-table”
Synset Definition

<synsetEntry id="s9439" author="122" version="29" timestamp="2014-11-17T07:49:46Z">
  <word ref="w9244">
    <definition source="ru.wiktionary" url="http://ru.wiktionary.org/wiki/cyp">
      Жидкое кушанье, обычно представляющее собой отвар с приправами и употребляемое как
    </definition>
    <example source="'Путешествие в седьмую сторону света', 2000, НКРЯ.">
      Был обед - овощной суп и курица на второе.
    </example>
  </word>
  <word ref="w40078"/>
  <word ref="w2893"/>
</synsetEntry>

{суп, бульон, похлёбка (soup)}
Current Problems

• Organizational issues.
  • The number of synsets was growing; moderators were not able to assess edits.

• Synset duplication.
  • Participants do not consult other people’s work.

• Hyponymy confusion.
  • In some cases, participants mix hyponymy with synonymy.

Genre → mechanized labor?
Evaluation: Size

- We compared YARN with other Russian thesauri.

<table>
<thead>
<tr>
<th></th>
<th># of concepts</th>
<th># of relations</th>
<th># of words</th>
<th>Availability</th>
<th>Commercial Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>RussNet</td>
<td>5.5K</td>
<td>8K</td>
<td>15K</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Russian Wordnet</td>
<td>157K</td>
<td>—</td>
<td>124K</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>RuThes</td>
<td>55K</td>
<td>210K</td>
<td>158K</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>RuThes-lite</td>
<td>26K</td>
<td>108K</td>
<td>115K</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>YARN</td>
<td>44K</td>
<td>0</td>
<td>48.6K</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The table is present in the paper.
Evaluation: Quality

• We took **200** most frequently edited synsets and assessed the quality of each synset.
  • Scale: Excellent, Satisfactory, Bad.

• Then, we aggregated the 800 answers using the MV strategy with pessimistic ties.
Evaluation: Quality

• Krippendorff’s alpha $\alpha = 0.202$ due to the skewness of the answers.

• Given these results, we treat the top 200 synsets as sufficiently good.

<table>
<thead>
<tr>
<th>Table 2: YARN synset quality.</th>
<th>MV</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>103</td>
<td>37</td>
<td>62</td>
<td>21</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>70</td>
<td>3</td>
<td>43</td>
<td>11</td>
</tr>
<tr>
<td>Bad</td>
<td>27</td>
<td>0</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>40</td>
<td>117</td>
<td>43</td>
</tr>
</tbody>
</table>
Conclusion

• The deliverables of **YARN** are available on its website under **CC BY-SA** in **XML**, **CSV**, **RDF**.

  • https://russiananword.net/en/
  • https://nlpub.ru/YARN
  • https://github.com/russianwordnet
Future Plans

• Creating verb and adjective synsets.
• Establishing more relations.
• Development of automatic methods for quality assurance.
• Widening the audience.
• Development of crowd management techniques.
Thanks!

Dmitry Ustalov,
Ural Federal University.
• https://ustalov.name/en/
• dmitry.ustalov@urfu.ru

The present work is supported by the Russian Foundation for the Humanities, project № 13-04-12020, and by the Mikhail Prokhorov Foundation.