Eliminating Fuzzy Duplicates in Crowdsourced Lexical Resources

Y. Kiselev, D. Ustalov, S. Porshnev
Ural Federal University
Yekaterinburg, Russia
Outline

• Introduction
• Related Work
• Problem
• Approach
• Experiments
• Results
• Discussion
• Conclusion
Introduction

• A good language resource should not include duplicated lexical senses.

• However, collaborative lexicography projects suffer from this problem.
  • Wiktionary, Yet Another RussNet, etc.

• We would like to address this problem.
Related Work

• **Automatic methods.**
  - Ontologies (Guarino & Welty, 2009),
  - Lexical resources (Sagot & Fišer, 2012).

• **Crowdsourcing methods.**
  - Find-Fix-Verify (Bernstein et al., 2010),
  - LR enrichment (Sajous et al., 2013).
Problem

• We focus on the synsets represented in WordNet-like thesauri.

• Example from the Russian Wiktionary:
  1) {стоматолог (stomatologist), дантист (dentist), зубной врач (“tooth doctor”)},
  2) {дантист (dentist), стоматолог (stomatologist)}.

• Expert-created LRs do not suffer.
Problem

• For the given example, the synset (2) is a subset of (1).

• Two problems:
  • to detect candidate synset pairs,
  • to confirm whether the synsets are duplicates, or not.
Approach

- Inspired by explanatory dictionaries.
  - Suppose the word $w$ has several meanings.
  - It is usually sufficient to provide one synonym for every sense of $w$.
  - A native speaker will be able to distinguish the meanings from each other.
Approach: Formulation

• Given a pair of different synsets $s_1$ and $s_2$, we treat them as duplicates if they share at least two words.

\[ \exists s_1 \in S, s_2 \in S : s_1 \neq s_2 \land |s_1 \cap s_2| \geq 2. \]

• This is a strong criterion that might be violated.
Approach: Two Stages

• **Filtering**, when possible duplicates are retrieved using the present criterion for further validation.

• **Voting**, when the obtained synset pairs are subject to manual verification.

• Our interest is to invite crowd workers to *refine* the crowd-created data.
Experiments

• Most crowdsourcing platforms are either not available or have insufficient number of Russian speakers.
  • Mechanical Turk,
  • CrowdFlower,
  • Prolific Academic, etc.

• The volunteers have been invited from VK, Facebook and Twitter.
Experiments: Engine

• **Mechanical Tsar** is an open source crowd-sourcing engine.

• **Our configuration:** fixed # answers, majority voting, no worker ranking.

http://mtsar.nlpub.org/
Experiments: “Filtering”

- Two lexical resources:
  - Yet Another RussNet (crowdsourced),
  - RuThes-lite (expert-created).

- We retrieved 210 presumably duplicated synsets from each one:
  - 70 synsets have two words in common,
  - 70 synsets have three,
  - 70 have four.
Experiments: “Voting”

• The workers are confirming whether the synsets are duplicates, or not.
Results

• We used a gold standard derived from the Babenko dictionary by an expert lexicographer.

• Quality metrics: precision, recall, $F_1$.

\[
P(s) = \frac{|s \cap \mathcal{L}(s)|}{|s|}
\]

\[
R(s) = \frac{|s \cap \mathcal{L}(s)|}{|\mathcal{L}(s)|}
\]

Table 1: Synset quality.

|          | Avg P | Avg R | Avg $F_1$ |
|----------|-------|-------|-----------|
| BAB      | 1.000 | 0.661 | 0.796     |
| YARN, aligned | 0.901 | 0.634 | 0.744     |
| YARN, machine | 0.840 | 0.774 | 0.805     |
Results: Deduplication

- **YARN** $F_1$: 0.744 $\rightarrow$ 0.805.

| Table 2: Crowdsourcing synset deduplication. |  |  |
|---|---|---|
| # of common words | 2 | 3 | 4+ |
| YARN | 61/70 | 64/70 | 68/70 |
| RuThes-lite | 25/70 | 40/70 | 51/70 |

| Table 3: YARN synset deduplication. | Avg P | Avg R | Avg $F_1$ |
|---|---|---|---|
| YARN, *machine* | 0.840 | 0.774 | 0.805 |
| YARN, *crowd* | 0.852 | 0.764 | 0.805 |
Discussion: Ambiguity

• In some cases, a couple of synonyms is not sufficient to derive the meaning.
  • “woman thought to have evil magic powers”, “a woman who uses magic or sorcery”.
  • “a bed with a back”, “a bed without a back”.

• We suggest including definitions for vague concepts into wordnets.
Discussion: Pairwise

• Pairwise annotation was especially hard for the workers.
• The complexity is $O(|s_1|+|s_2|)$, e.g. $O(4+4)=8$ operations per pair.
• Task clustering and visual hints could be useful.

Table 4: Average synset sizes.

| # of common words | 2   | 3   | 4+  |
|-------------------|-----|-----|-----|
| YARN              | 4.2 | 4.6 | 5.5 |
| RuThes-lite       | 4.3 | 5.0 | 5.8 |
Discussion: Agreement

• The workers agreement did not change for any number of common words in an expert-created resource.

Table 5: # of merge decisions made unanimously.

| # of common words | 2       | 3       | 4+      |
|-------------------|---------|---------|---------|
| YARN              | 32/70   | 47/70   | 57/70   |
| RuThes-lite       | 36/70   | 35/70   | 32/70   |
Conclusion

• We found this approach useful for a crowdsourced resource even without “Voting” 😊.
  • But the Voting stage is useful for QA in expert-created resources.

• The results are published (CC BY-SA).
  • http://ustalov.imm.uran.ru/pub_duplicates-gwc.tar.gz.
Thanks!

Dmitry Ustalov,
Ural Federal University.

- [https://ustalov.name/en/](https://ustalov.name/en/)
- [dmitry.ustalov@urfu.ru](mailto: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.