



**Ural Federal
University**

named after the first President
of Russia B.N.Yeltsin

Expanding Hierarchical Contexts for Constructing a Semantic Word Network

Dmitry Ustalov

IMM UB RAS / UrFU
Yekaterinburg, Russia

Outline

- Introduction
- Related Work
- The WATLINK Method
- Evaluation
- Lexical Relations from the WotC
- Conclusion

Introduction

“Hypernymy extraction {...} may be the best thing since sliced bread.”

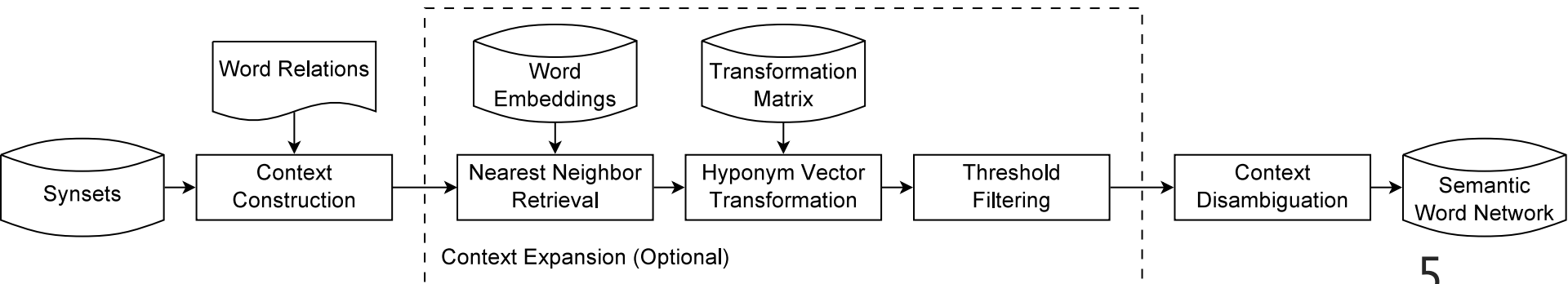
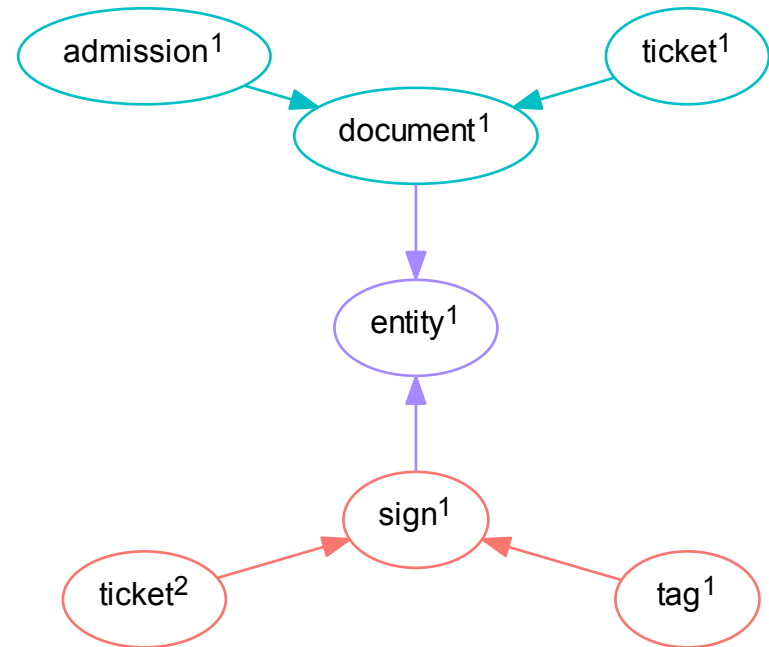
- Noisy hypernymy dictionaries:
 - *is-a* dictionaries,
 - corpora,
 - crowdsourcing, etc.
- **The goal** is to *propagate* hypernyms to the less-covered words.

Related Work

- **Lexicographers:** WordNet, RuThes, etc.
- **Crowdsourcing:** Wiktionary, etc.
- **Sem. Classes:** Pantel & Lin (2000s).
- **Matching + MT:** BabelNet, etc.
- **Projection Learning:** Fu et al. (2014).
- **NNs:** HypeNET, LexNET, etc.

WATLINK: Overview

- **Input:** a set of synsets S , a set of relations R .
- **Output:** a graph N that connects word senses via hypernymy.



WATLINK: Construction

- For each synset, a **hierarchical context** is constructed.
 - Such a context represents all the hypernyms for each word.
- e.g.** $\text{hctx}(\{auto^2, car^1, automobile^1\})$ is $\{vehicle, transport, motor\ vehicle\}$.
- Some words are more important, some are less important → use **tf-idf**.

WATLINK: Disambiguation

- For each word in hierarchical context, the sense label is estimated.

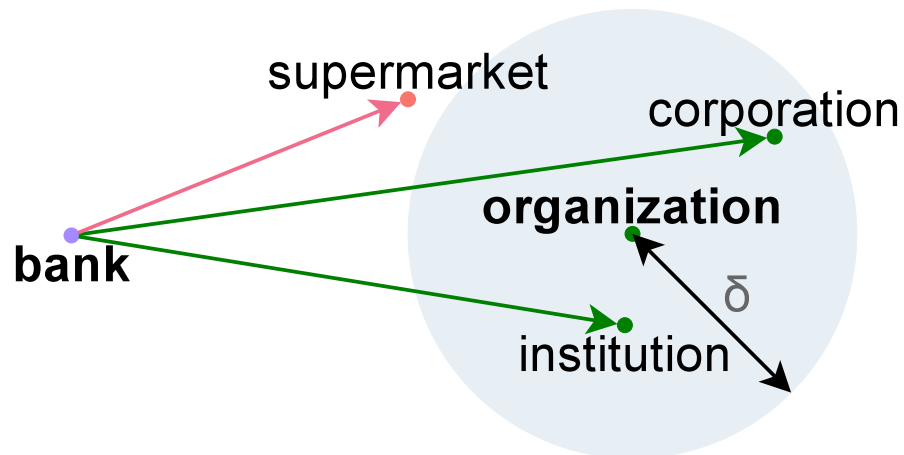
$$\hat{h} = \arg \max_{\substack{S' \in S, S \neq S', h' \in S', \\ words(\{h'\}) = \{h\}}} cos(hctx(S), S')$$

e.g. *material*, *data* is likely to be similar to $\{information^1, data^1, material^1\}$ rather than to $\{material^2, textile^1\}$.

- Then, each synset is multiplied to its disambiguated context \rightarrow graph **N**.

WATLINK: Expansion (Optional)

- Extract n nearest neighbors of each hypernym using w2v.
- Multiply each hyponym vector to the “projection” matrix and ensure that the resulting vector is within δ .



Evaluation: Setup

- The **synsets** are obtained using the WATSET method (our ACL 2017 paper).
 - 55 369 synsets uniting 83 092 words
- **Hypernyms** from patterns, Wiktionary, SAD → Joint (150K pairs w/o Exp).
- **Projection learning**: 20 matrices, 500 dimensions (our EACL 2017 paper).
- **Meta-parameters**: $n=10$, $\delta=0.6$.

Evaluation: Data & Measures

- **Measures:** precision, recall, F_1 -score.
- **RuWordNet:** a pair is matched iff there is a path from hyponym to hypernym.
- **LRWC:** “yes” or “no” using microtask-based crowdsourcing.

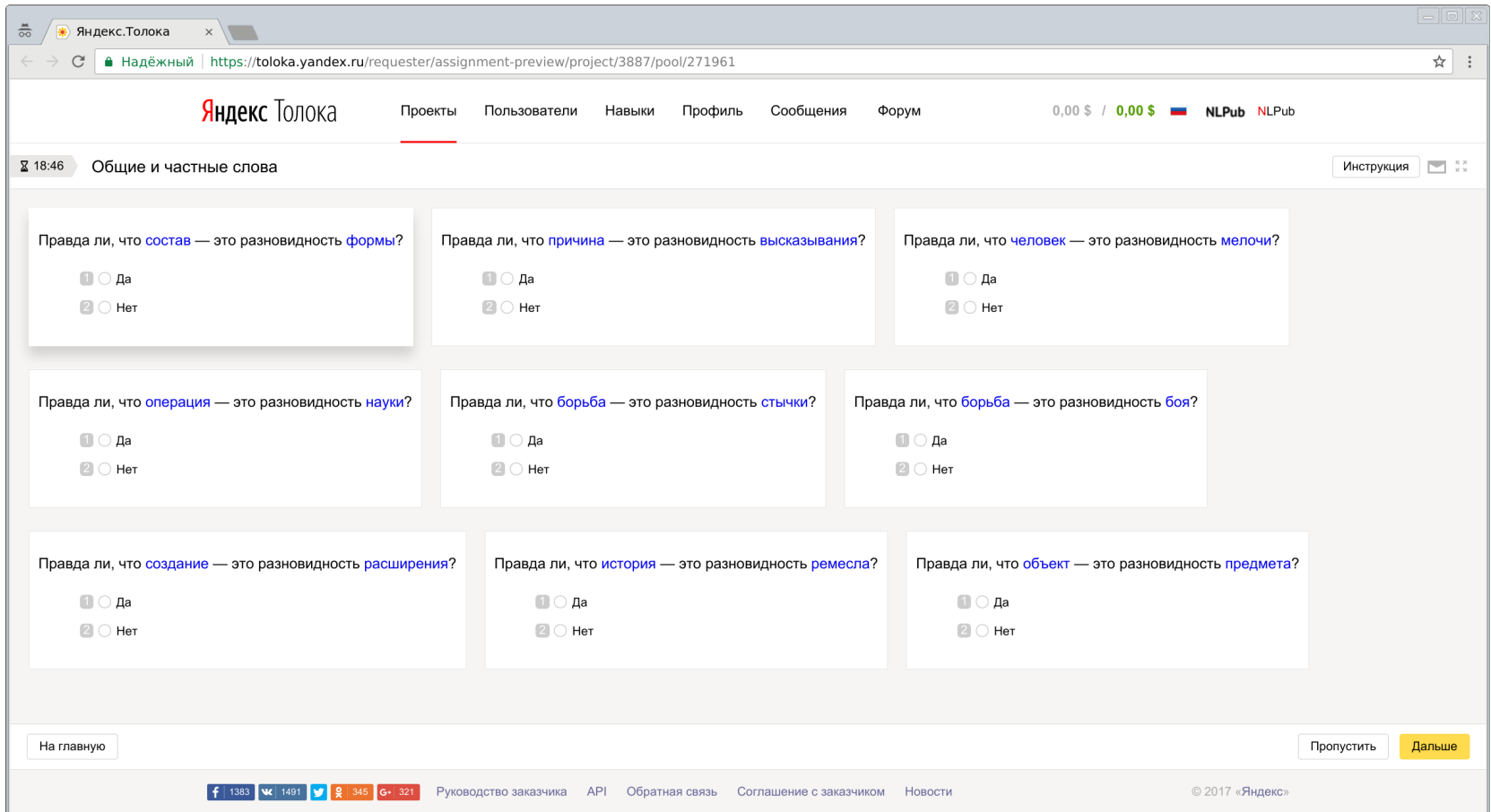
Evaluation: RuWordNet

Method	# of pairs	Precision	Recall	F ₁ -score
Patterns	1,597,651	0.1611	0.3255	0.2155
Patterns + SWN	236,922	0.1126	0.2451	0.1543
Patterns + Limit	10,458	0.3773	0.0157	0.0302
Patterns + Limit + Exp	10,715	0.3760	0.0160	0.0307
Patterns + Limit + SWN	46,758	0.1140	0.0717	0.0880
Patterns + Limit + Exp + SWN	47,387	0.1129	0.0722	0.0881
Wiktionary	108,985	<u>0.3877</u>	0.0898	0.1458
Wiktionary + Exp	110,329	0.3874	0.0907	0.1469
Wiktionary + SWN	177,787	0.1836	0.3460	0.2399
Wiktionary + Exp + SWN	179,623	0.1844	0.3464	<u>0.2407</u>
SAD	36,800	0.1823	0.1502	0.1647
SAD + Exp	37,702	0.1825	0.1515	0.1655
SAD + SWN	99,678	0.1385	0.1883	0.1596
SAD + Exp + SWN	98,085	0.1383	0.1879	0.1593
Joint	149,195	0.1719	0.2590	0.2067
Joint + Exp	151,150	0.1720	0.2594	0.2069
Joint + SWN	218,290	0.1687	<u>0.3867</u>	0.2350
Joint + Exp + SWN	216,285	0.1685	0.3865	0.2347

Lexical Relations from the WotC

- Lexical Relations from the Wisdom of the Crowd (**LRWC**).
 - **License:** CC BY-SA 3.0.
 - <https://nlpub.ru/LRWC>
- 300 most frequent Russian nouns.
 - 10 600 annotated unique pairs.
 - 7 annotators per task, strict control.
- **DOI:** [10.5281/zenodo.546302](https://doi.org/10.5281/zenodo.546302)

Yandex.Toloka: Example



The screenshot shows the Yandex.Toloka interface for a task assignment preview. The browser address bar shows the URL: <https://toloka.yandex.ru/requester/assignment-preview/project/3887/pool/271961>. The page header includes the Yandex.Toloka logo, navigation links (Проекты, Пользователи, Навыки, Профиль, Сообщения, Форум), and a balance display (0,00 \$ / 0,00 \$). The main content area is titled "Общие и частные слова" and contains nine questions arranged in a 3x3 grid. Each question asks if a word is a synonym of another, with two radio button options: "Да" (Yes) and "Нет" (No).

Question	Options
Правда ли, что состав — это разновидность формы ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что причина — это разновидность высказывания ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что человек — это разновидность мелочи ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что операция — это разновидность науки ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что борьба — это разновидность стычки ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что борьба — это разновидность боя ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что создание — это разновидность расширения ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что история — это разновидность ремесла ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет
Правда ли, что объект — это разновидность предмета ?	1 <input type="radio"/> Да 2 <input type="radio"/> Нет

At the bottom of the page, there are buttons for "На главную" (Back to home), "Пропустить" (Skip), and "Далее" (Next). The footer includes social media links (Facebook, VK, Twitter, Instagram, Google+) and a copyright notice: © 2017 «Яндекс».

Quality Control: Surprise!!

3bfcca05aac7939942a9107830b46212

10 минут назад

Общие и частные слова — 2017-04-14



здравствуйте хорошее задание. когда ешо появятся задания?

Я

5 минут назад

Спасибо.

3bfcca05aac7939942a9107830b46212

минуту назад

когда ешо появятся?

Я

несколько секунд назад

Всегда вам рады.

Evaluation: LRWC

Method	Precision	Recall	F ₁ -score
RuThes	0.7035	0.9168	<u>0.7961</u>
Joint + Exp	0.6719	0.9002	0.7695
Joint	0.6726	0.8975	0.7690
Wiktionary + SWN	0.6287	0.8775	0.7326
Wiktionary + Exp + SWN	0.6254	0.8779	0.7304
Joint + SWN	0.5590	<u>0.9306</u>	0.6985
Joint + Exp + SWN	0.5569	0.9304	0.6968
RWN (Nouns)	0.5878	0.8400	0.6917
SAD + Exp	0.6313	0.6141	0.6226
SAD	0.6321	0.6121	0.6220
Patterns	0.4821	0.8710	0.6207
Wiktionary + Exp	0.7488	0.3485	0.4756
Wiktionary	<u>0.7492</u>	0.3467	0.4741
Patterns + Limit	0.6711	0.3103	0.4244
Patterns + Limit + Exp	0.6700	0.3105	0.4244

Conclusion

- WATLINK propagates and disambiguates hypernyms.
 - It improves the hypernymy extraction recall.
 - This is useful for processing not-so-frequent words.
- GitHub: [dustalov/watlink](https://github.com/dustalov/watlink), [nlpub/hyperstar](https://github.com/nlpub/hyperstar).
- LRWC 1.1: [10.5281/zenodo.546302](https://doi.org/10.5281/zenodo.546302).

Thanks!

Dmitry Ustalov,
IMM UB RAS / UrFU.

- <https://nlpub.ru/>
- dmitry.ustalov@urfu.ru



The reported study is funded by **RFBR** according to grant no. 16-37-00354 мол_a. The author grateful to Microsoft Research for providing free access to computational resources of the **Microsoft Azure** cloud under the Azure for Research Award program.