RuSemCor: A Word Sense Disambiguation Corpus for Russian

Alexander Kirillovich, Ilia Karpov, Natalia Loukachevitch, Maksim Kulaev, Dmitry Ilvovsky

HSE University
Moscow, Russia
dilv_ru@yahoo.com

Summary

We present RuSemCor, an open Word Sense Disambiguation (WSD) corpus for Russian. The corpus was constructed by manually linking tokens from the OpenCorpora corpus to senses in the Russian wordnet RuWordNet. It consists of 869 documents with 121,710 tokens of which 51,588 are wordnet annotated. The resource is represented using the NIF, OLiA, OntoLex, and Global WordNet ontologies and integrated into the Linguistic Linked Open Data cloud.

We used RuSemCor as a diagnostic benchmark to evaluate a range of WSD methods. Our experiments yielded three main findings. 1) Generative LLMs substantially outperform traditional knowledge-based methods such as Personalized PageRank. 2) Despite their strengths, generative LLMs do not surpass encoder-based models specifically trained for WSD. 3) Incorporating lexical-semantic relations from RuWordNet produces mixed results: it enhances the performance of encoder-based models and leading LLMs like GPT- 4, DeepSeek, and Mistral 24B, but tends to degrade accuracy for smaller generative models such as GPT-3 and Mistral 7B.
The resource is distributed under the CC BY-SA open license and is available at:
http://github.com/llod-ru/rusemcor

RuSemCor statistics

Entity	Count
Documents	869
Sentences	7,398
Tokens	
All tokens	121,710
Sense-annotated tokens	51,588
Lemmas	
All lemmas	18,244
Sense-annotated lemmas	10,680
Covered RWN synsets	9,075
Covered RWN lexical entries	9,211
Covered RWN senses	11,152

Corpus description

RuSemCor is a corpus of Russian texts with morphological and lexical-semantic annotations, built from OpenCorpora by manually linking its tokens to senses in RuWordNet.

Source Resources:

- **OpenCorpora**: An open Russian corpus with morphological annotations (lemmas, POS tags, grammatical features).
- **RuWordNet (RWN):** A Russian wordnet with synsets for nouns, verbs, and adjectives.

Corpus construction

1. **Automatic Task Generation**: For each token, a list of candidate RWN synsets was generated by aligning via lemma and POS. This required resolving differences in POS inventories and lemmatization between the two resources.
2. **Sampling**: 869 mid-sized documents were manually sampled for annotation.
3. **Manual Annotation**: Three trained annotators, supervised by an expert, selected the correct synset for each token. High inter-annotator agreement (96% F1) was achieved. *Note: Not all tokens are linked to RWN, due to: missing POS categories, missing lexical entries, or no suitable synset in context.*
4. **Final Assembly**: The completed annotations were processed into the final corpus, combining sense annotations from RWN with text metadata and morphology from OpenCorpora.

Experimental setups (WSD)

To investigate whether other structured information from the wordnet graph — especially semantic relations — can improve disambiguation accuracy, we designed and evaluated several model setups. Each setup constructs the sense description from different combinations of synset features derived from RuWordNet. **These features are**: artificial numerical ID (id), canonical name (name), definition (def), lexical entries (entries), hyponyms (hypos), hypernyms (hypers), and the hyponym–hypernym chain from the synset to the root of the graph, represented both as se- quences of synset names (chain) and of numerical IDs (idchain).
For this evaluation, we adapted **ConSeC** to Russian by employing the pre-trained model deepvk/deberta-v1-base as the encoder backbone. This model is a Russian-language DeBERTa-v1-base variant pre-trained on 400 GB of diverse Russian texts, including Wikipedia, books, social media, and news. With this backbone we fine-tuned ConSeC on the training subset of our RuSemCor corpus.

Prompt for LLMs

“Consider the following sentence: «Право на жизни находится под угрозой во всем мире, говорит Папа.». In what sense the word «Папа» is used in this sentence:
1) The sense «ОТЕЦ», that can be expressed by the following terms: БАТЯ, ПАПА, ПАПУЛЯ, БАТЮШКА, ПАПОЧКА, ОТЕЦ, ПАПАША, БАТЬКА, ПАПЕНЬКА. The hypernyms of this sense are the following: РОДСТВЕННИК-МУЖЧИНА, РОДИТЕЛЬ. The hyponyms of this sense are the following: ПРАДЕД, ОТЕЦ МУЖА ИЛИ ЖЕНЫ.
2) The sense «ПАПА РИМСКИЙ», that can be expressed by the following terms: ПАПА РИМСКИЙ, ПАПА, ПОНТИФИК, РИМСКИЙ ПАПА, СВЯТЕЙШЕСТВО, СВЯТОЙ ПАПА, ПАПСТВО, ЕГО СВЯТЕЙШЕСТВО, СВЯТЕЙШИЙ ПАПА, ГЛАВА КАТОЛИЧЕСКОЙ ЦЕРКВИ. The hypernyms of this sense are the following: ГЛАВА ГОСУДАРСТВА, ГЛАВА ЦЕРКВИ, КАТОЛИЧЕСКИЙ СВЯЩЕННИК. The definition of this sense is «верховный глава католической церкви; папа».
Please, print only number of the appropriate sense and nothing else.”

Results of evaluation of WSD methods

Method	Accuracy (%)
Baseline	
Random	63.60
Min id	70.91
Most frequent	74.10
Most frequent + min id	84.54
Knowledge-based	
Personalized PageRank	74.30
Finetuned encoder	
ConSeC	
id	87.88
name	91.25
chain	92.57
idchain	92.20
name+def	92.42
name+def+entries+hypos+hypers	**92.70**
Generative LLM	
GPT-3.5	
name+def	83.95
name+def+entries+hypos+hypers	82.92
Mistral 7B Instruct v0.3	
name+def	83.77
name+def+entries+hypos+hypers	83.18
name+def+entries+hypos+hypers+chain	82.06
Mistral Small 24B Instruct 2501	
name	89.52
def	78.42
entries	85.32
hypos	80.77
hypers	83.07
hypos+hypers	86.57
chain	88.12
name+def	89.21
name+def+entries	90.15
name+def+hypos+hypers	90.88
name+def+chain	90.05
name+def+entries+hypos+hypers	90.80
name+def+entries+hypos+hypers+chain	90.63
Deepseek	
name	91.50
def	76.86
entries	86.50
hypos	81.86
hypers	82.75
hypos+hypers	87.87
chain	90.21
name+def	91.15
name+def+entries	91.41
name+def+hypos+hypers	91.99
name+def+chain	91.43
name+def+entries+hypos+hypers	91.96
name+def+entries+hypos+hypers+chain	91.88
GPT-4	
name+def	91.56
name+def+entries+hypos+hypers	92.09

In Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM ’25), November 10–14, 2025, Seoul, Republic of Korea. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3746252.3761624