Taxonomy Enrichment

Skoltech

Alexander Panchenko
A.Panchenko@skoltech.ru

Acknowledgements

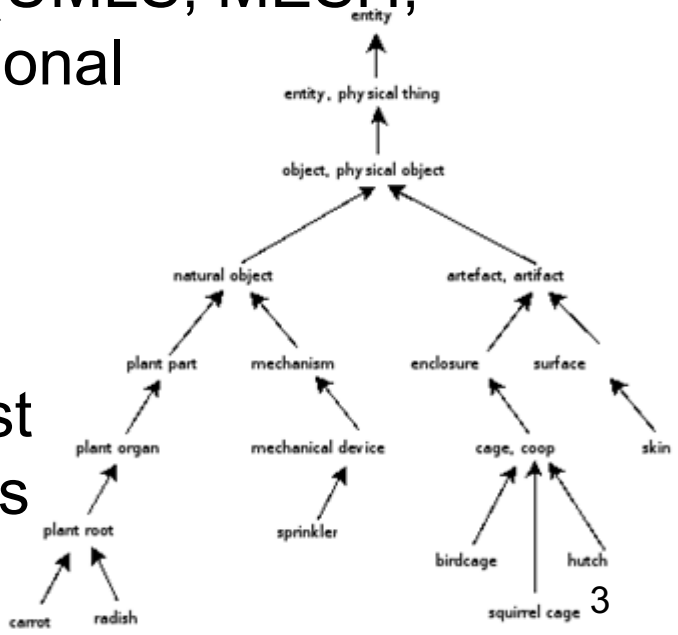
- Based on the following publications:
  1. Nikishina, I., Logacheva, V., Panchenko, A., and Loukachevitch N. (2020): **Studying Taxonomy Enrichment on Diachronic WordNet Versions**. In Proceedings of the 28th International Conference on Computational Linguistics (COLING-2020). Barcelona, Spain.
  2. Nikishina, I., Logacheva, V., Panchenko, A., and Loukashevich N. (2020): RUSSE-2020: **Findings of the First Taxonomy Enrichment Task for the Russian Language**. In Proceedings of the 26th International Conference on Computational Linguistics and Intellectual Technologies (Dialogue'2020). Moscow, Russia. RGGU
  3. Nikishina, I., Logacheva, V., Panchenko, A., and Loukashevich N. (2021): **Exploring Graph-based Representations for Taxonomy Enrichment**. Global WordNet Conference. Pretoria. South Africa
- This presentation is based on the materials by Natalia Loukachevitch for the Dialogue Evaluation 2020 [2].

Lexical and domain knowledge in NLP

- Lexical relations:
  - WordNet and wordnets for different languages
  - ImageNet was constructed over WordNet

- Domain knowledge
  - Medical ontologies and thesauri (UMLS, MESH, Gene Ontology) are very influentional in medical NLP and bioNLP

- Necessity of large resources
  - Taxonomy is a back-bone of most knowledge representation models



Taxonomy relations

- Important for:
  - lexical inference, question-answering, query expansion
- Can have different names in specific resources
  - Hyponym-hypernym relations in lexical-semantic resources
  - Class-subclass relations in ontological resources, etc.
- Problem: any resources are never complete

- Why not just use word embeddings?
  - Modern word embeddings can capture word relatedness
  - But they mix all types of semantic relations together
  - No guaranties of any real explainable relation

Russian examples: most similar words according to news collection

- **моторшоу** : автошоу 0.861 ; автовыставка 0.821 ; автомобильный_салон 0.778 ; автосалон 0.704…
- **месседж** : посыл 0.837 ; тезис 0.623 ; ремарка 0.540 ; постулат 0.538 ; клише 0.536…
- **сдавление** : сдавливание 0.805 ; спинной 0.682 ; омертвение 0.626 ; опущение 0.626..
- **поэтичность** : проникновенность 0.795 ; образность 0.738 ; живость 0.7353 ..
- **сифон** : взбивание 0.616 ; баллон 0.608 ; микроволновка 0.604 ; дозатор 0.602..
- **братание** : братоубийство 0.571 ; монархизм 0.558 ; коллаборационизм 0.557 ;

English examples: most similar words according to JoBimText count-based model

Jos

Jo	Score
mouse#NN	746
rat#NN	192
rodent#NN	122
monkey#NN	112
pig#NN	103
animal#NN	95
human#NN	94
rabbit#NN	91
keyboard#NN	91
cow#NN	83
hamster#NN	82
frog#NN	81
cat#NN	80
bird#NN	79
squirrel#NN	72
snake#NN	70
sheep#NN	67
flies#NN	66
deer#NN	65
goat#NN	65
cattle#NN	63
mammal#NN	62
joystick#NN	61

Bims

Bim	Score	Count
click#NN#-prep_of	14433.61	
a#DT#det	11612.08	
click#NN#-nn	9071.84	
the#DT#det	8613.77	
keyboard#NN#-conj_and	7548.80	
cat#NN#-conj_and	5417.09	
computer#NN#nn	4776.27	
keyboard#NN#conj…	4241.33	
button#NN#-nn	3987.05	
pad#NN#-nn	3320.76	
rat#NN#conj_and	2971.02	
rat#NN#-conj_and	2821.09	
click#VB#-dobj	2472.84	
man#NN#conj_and	1958.96	
use#VB#-dobj	1859.87	
white-footed#JJ#amod	1810.81	
your#PRP$#poss	1791.63	
cell#NN#-nn	1694.21	
laboratory#NN#nn	1667.45	
normal#JJ#amod	1533.29	
cursor#NN#-nn	1409.18	
strain#NN#-prep_of	1352.15	

CW

Sense 0 168: rat#NN · rodent#NN · monkey#NN · pig#NN · animal…

Sense 1 32: keyboard#NN · joystick#NN · stylus#NN · printer#NN · …

http://ltmaggie.informatik.uni-hamburg.de/jobimviz/?

Distribution across the relation types: contextualized neural models

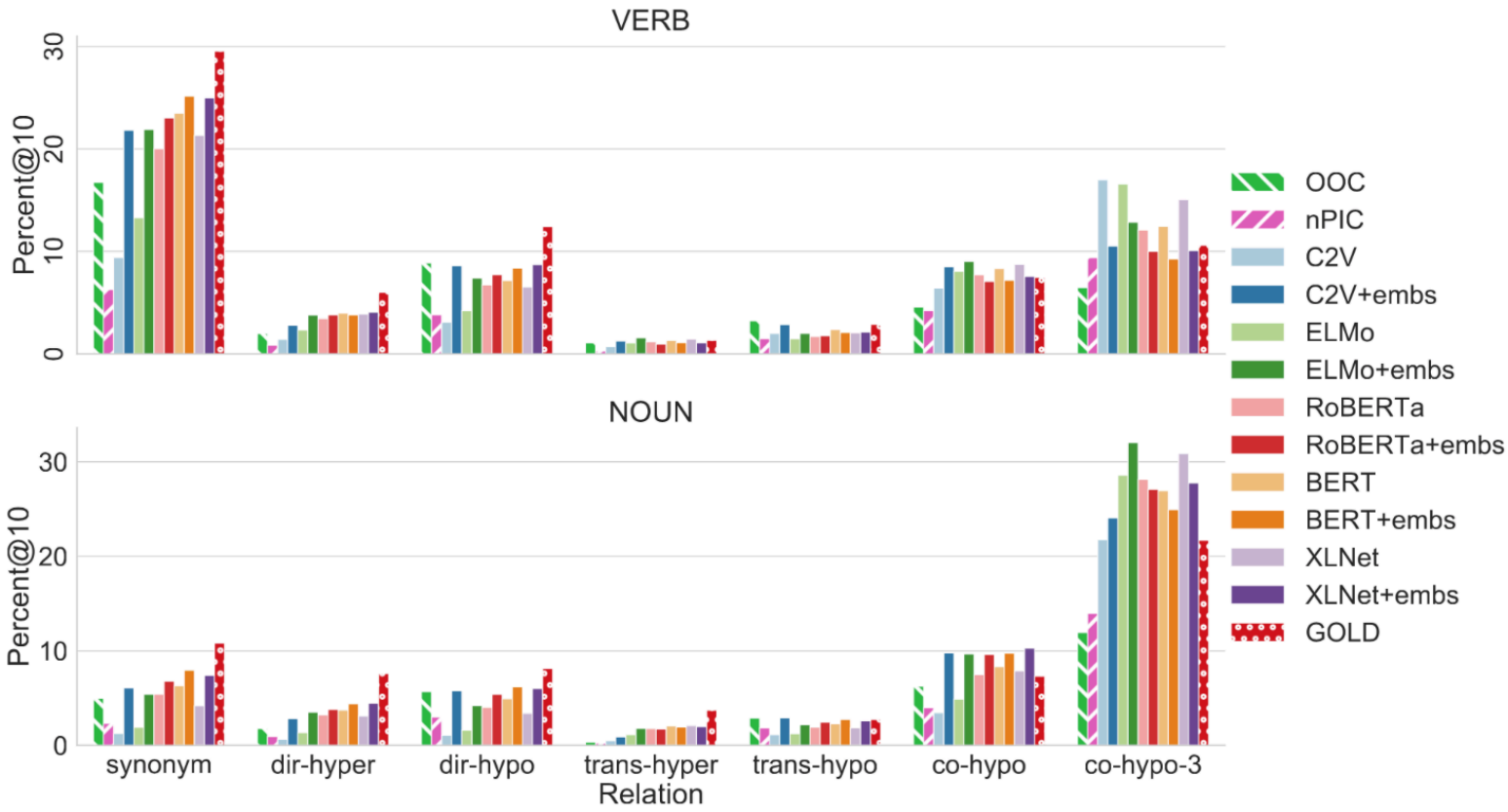


Figure 2: Proportions of substitutes related to the target by various semantic relations according to WordNet. We took top 10 substitutes from each model and all substitutes from the gold standard.

Arefyev, N., Sheludko, B., Podolskiy, A., and Panchenko A. (2020): Always Keep your Target in Mind: Studying Semantics and Improving Performance of Neural Lexical Substitution. In Proceedings of the 28th International Conference on Computational Linguistics (COLING-2020). Barcelona, Spain.

Distribution across the relation types: contextualized neural models

We were not able to travel in the weather , and there was no ***phone*** .										
GOLD	**telephone (5)**									
OOC	phone	**telephone**	phones	cellphone	fone	videophone	handset	telephones	p990i	cell-phone
XLNet	electricity	internet	phone	power	**telephone**	car	water	communication	radio	tv
XLNet+embs	phone	**telephone**	phones	cellphone	internet	radio	electricity	iphone	car	computer
What happened to the big , new garbage ***can*** at Church and Chambers Streets ?										
GOLD	**bin (4)**	**disposal (1)**	**container (1)**							
OOC	can	could	should	would	will	must	might	to	may	ll
XLNet	can	dump	**bin**	truck	**disposal**	pit	heap	pile	**container**	stand
XLNet+embs	can	could	will	**bin**	cannot	dump	may	truck	**disposal**	stand

Types of semantic relations: synonym, co-hyponym, co-hyponym 3, target, direct hypernym, transitive hypernym, direct hyponym, transitive hyponym, unknown-relation, unknown-word

Figure 3: Examples of top substitutes provided by annotators (GOLD), the baseline (OOC), and two presented models (XLNet and XLNet+embs). The target word in each sentence is in bold, true positives are in bold also. The weights of gold substitutes are given in brackets. Each substitute is colored according to its relation to the target word. Substitutes before post-processing are shown.

Arefyev, N., Sheludko, B., Podolskiy, A., and Panchenko A. (2020): Always Keep your Target in Mind: Studying Semantics and Improving Performance of Neural Lexical Substitution. In Proceedings of the 28th International Conference on Computational Linguistics (COLING-2020). Barcelona, Spain.

Methods for hypernym detection

- Lexical-semantic methods (Hearst's patterns)
  - X is a Y
  - Various modifications: syntactic-based, with SVD, etc.
- Distributional methods (embedding-based methods)
  - Unsupervised methods based on operations over embeddings
- Supervised methods
  - Machine learning methods such as SVM or neural networks
  - Projection-learning – learning linear transformations over clustered word embeddings
- Combined methods

Hypernym detection: datasets and evaluations

- Classification task
  - Datasets with comparable numbers of positive and negative examples for all types of relations
  - Measures: F-measure and Accuracy
  - But in reality the number of positive examples for any relation is much smaller
- SemEval-2016 Task 14 (taxonomy enrichment)
  - Systems should attach new word to WordNet having sense definition
- Ordering task: candidates should be ordered
  - SemEval-2018 Task 9 (hypernymy discovery),
  - Restricted corpus; Ordering measures: MAP, MRR

RUSSE'2020 evaluation

- Published RuWordNet – 110 thousand Russian words and expressions
- New version of RuWordNet – 130 thousand Russian words and expressions is prepared but not published
- An associate text corpus
- Evaluation task
  - For new words (noun and verbs) to predict the nearest synsets from the published version
  - Correct answers should indicate
    - Direct hypernyms if a new word created a new synset
    - Hypernyms of direct hypernyms

Preparing dataset for evaluation

- We extracted words (nouns and verbs), which are present in the extended RuWordNet, but absent in the published RuWordNet. From the list, the following words were excluded:
  - all three-symbol words and the majority of four-symbol words;
  - diminutive word forms and feminine gender-specific job titles;
  - words which are derived from words which are included in the published RuWordNet;
  - words denoting inhabitants of cities and countries;
  - geographic and personal names;
  - compound words that contain their hypernym as a substring.

Datasets

- The extracted words were subdivided into public and private tests.
- Besides, training set from RuWordNet lower level synsets was generated

	Nouns	Verbs
Total in RuWordNet	29 297	7 636
Train set	12 393	2 109
Private test set	1 525	350
Public test set	763	175

Examples of words to add (orphans)

- абдоминопластика
- абсентеизм
- абсолютизация
- абсорбент
- абстракционизм
- абстракционист
- аваль
- аванзал
- аварийщик
- автаркия
- авуары
- агитпункт
- агностик
- адвентист
- адгезия
- аджика
- адсорбент
- адъюнкт
- адъюнктура
- азовка
- азу
- айпад
- айран
- абсолютизировать
- адсорбировать
- акать
- активировать
- актировать
- алеть
- американизировать
- аннигилировать
- аннотировать
- анодировать
- аукаться
- бередить
- бинтовать
- бичевать
- блокироваться
- бодриться
- бряцать
- булькать
- буреть
- бутилировать
- вальсировать
- вбухивать

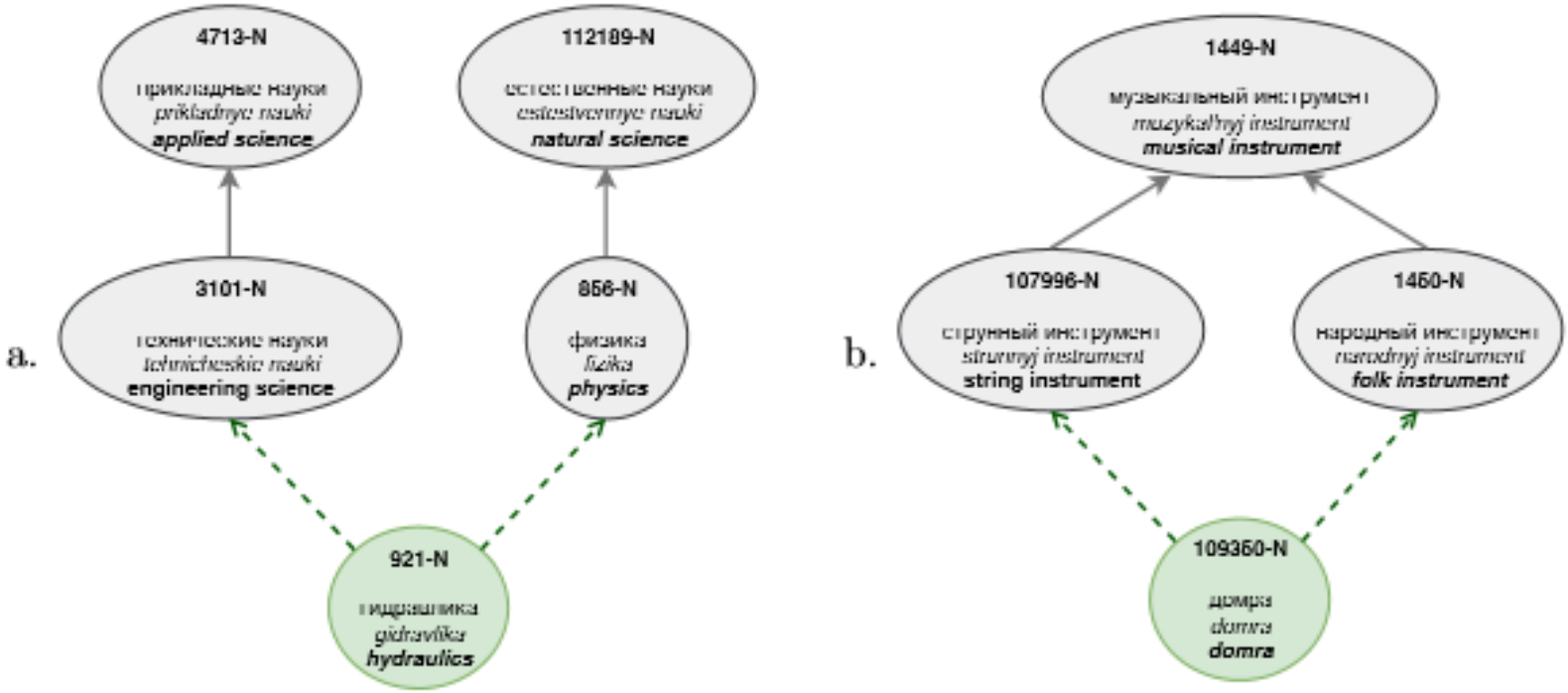
English dataset: WordNet versions

Taxonomy	Synsets		Lemmas		New words	
	Nouns	Verbs	Nouns	Verbs	Nouns	Verbs
WordNet 1.6	66 025	12 127	94 474	10 319	-	-
WordNet 1.7	75 804	13 214	109 195	11 088	11 551	401
WordNet 2.0	79 689	13 508	114 648	11 306	4 036	182
WordNet 2.1	81 426	13 650	117 097	11 488	2 023	158
WordNet 3.0	82 115	13 767	117 798	11 529	678	33

Dataset	Nouns	Verbs
WordNet 1.6 - WordNet 3.0	17 043	755
WordNet 1.7 - WordNet 3.0	6 161	362
WordNet 2.0 - WordNet 3.0	2 620	193
RuWordNet 1.0 - RuWordNet 2.0	14 660	2 154
RUSSE'2020	2 288	525

Procedure of evaluation

- Ordering task, measures: MAP
- Systems have to find relatively precise place for adding orphans: gold standard hypernym, or its hypernym
  - There can be several hypernyms on each level
  - Calculation of connectivity components



Baseline approach

- Hypernyms of fastText nearest neighbours
  - Represent each synset as a sum of its lemma embedings.
  - Get top k = 10 closest synsets the input word (~co-hyponyms).
  - Output 10 most frequent hypernyms of the “co-hyponymy synsets”.

Baseline approach: wny it works?

http://ltmaggie.informatik.uni-hamburg.de/jobimviz/

CW

Show all senses

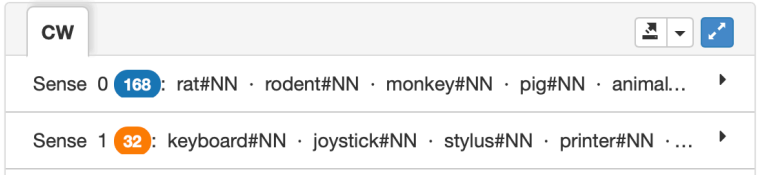
Sense Terms	IS-As
rat#NN	animal:315301
rodent#NN	specy:58437
monkey#NN	wildlife:25284
pig#NN	mammal:14342
animal#NN	part:12802
human#NN	predator:11700
rabbit#NN	food:10868
cow#NN	of animal:10710
hamster#NN	problem:10626
frog#NN	creature:10318
cat#NN	
bird#NN	
squirrel#NN	
snake#NN	
sheep#NN	
flies#NN	
goat#NN	
deer#NN	
cattle#NN	
mammal#NN	
mosquito#NN	
lizard#NN	
primate#NN	
dog#NN	
ferret#NN	

CW

Show all senses

Sense Terms	IS-As
keyboard#NN	device:30576
joystick#NN	product:12672
stylus#NN	item:6750
printer#NN	equipment:4730
modem#NN	technology:4191
monitor#NN	tool:3750
peripheral#NN	bit:2150
scanner#NN	way:2128
remote#NN	system:2040
laptop#NN	electronic:1751
headset#NN	
keypad#NN	
headphones#NN	
trackpad#NN	
computer#NN	
button#NN	
cursor#NN	
pc#NN	
projector#NN	
camera#NN	
webcam#NN	
controller#NN	
interface#NN	
microphone#NN	
phone#NN	

Two senses of the word “mouse”:



Baseline improvements

- **Ranking.** Ranking Extended Hypernyms List by Weighted Similarity
  - Consdier hypernyms of hypernyms as well to pull in more candidates
  - We assume that the most frequent and the most similarcandidates are the true hypernyms of the wo
  - Count the most frequent synset and rank

$$score_{h_i} = n \cdot sim(v_o, v_{h_i})$$

- **Wiki.** Features Extracted from Wiktionary
  - the candidate is present in the Wiktionary hypernyms list for the input word (binary feature)
  - the candidate is present in the Wiktionary synonyms list (binary feature)
  - the candidate is present in the Wiktionary definition (binary feature)
  - average cosine similarity between the candidate and the Wiktionary hypernyms of the input word
  - Supervised combination of all features using logistic regression.

The best approach (Yuriy)

- Candidates were ranked by a linear model with handcrafted weights. The list of features includes:
  - top 10 similar words from WordNet, their hypernyms and hypernyms of hypernyms;
  - hypernyms or hypernyms of hypernyms on Wiktionary page;
  - “en-ru” translation of WordNet hypernyms of “ru-en” translation of the word (extracted with Yandex Machine Translation model);
  - candidate is in the word definition in the Wiktionary page;
  - candidate is in the Yandex or Google search result pages.

Evaluation results: WordNet and RuWordNet

Method	English		Russian	
	Nouns	Verbs	Nouns	Verbs
	fastText			
Baseline	0.325	0.183	0.421	0.334
Ranking	0.375	0.190	0.507	0.336
Ranking + Wiki	**0.400**	**0.238**	0.540	0.383
	BERT			
Baseline	0.239	0.097	0.138	0.119
Ranking	0.238	0.105	0.185	0.127
Ranking + Wiki	0.253	0.120	0.218	0.161
	RUSSE'2020 participating systems			
Top-1 for Nouns: *Yuriy*	0.328	0.230	**0.552**	0.436
Top-1 for Nouns: *Yuriy*, no search engine features	0.300	0.231	0.507	0.388
Top-1 for Verbs: (Dale, 2020)	0.234	0.224	0.418	**0.448**

Table 3: MAP scores for the taxonomy enrichment methods for English (2.0-3.0) and Russian datasets.

Evaluation results: different word categories

Method	Nouns				Verbs		
	NE	Short	Other	All	Short	Other	All
% in the data	38%	6%	61%	–	5%	95%	–
	WordNet 2.0 — WordNet 3.0 (fastText)						
Baseline	0.328	0.319	0.233	0.291	**0.444**	0.191	0.205
Ranking	0.424	0.348	0.288	0.339	0.296	0.208	0.213
Ranking + Wiki	**0.437**	**0.360**	**0.332**	**0.372**	0.411	**0.263**	**0.271**
	RuWordNet 1.0 — RuWordNet 2.0 (fastText)						
% in the data	25%	7%	70.7%	–	0%	100%	–
Baseline	0.251	0.165	0.337	0.309	-	0.232	0.232
Ranking	0.417	0.218	0.381	0.384	-	0.252	0.252
Ranking + Wiki	**0.436**	**0.230**	**0.416**	**0.414**	-	**0.295**	**0.295**

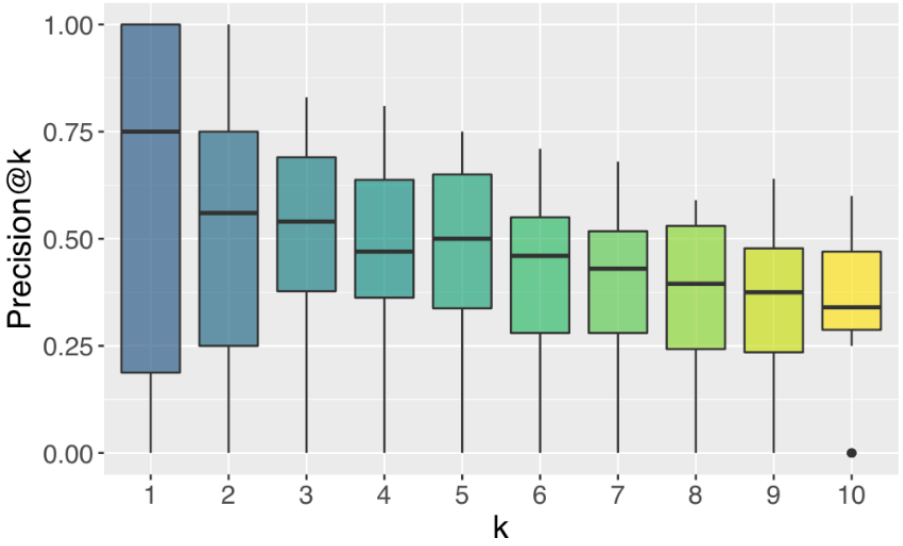
Comparison of predictions with gold standard

Rank	dancing-master	(to) ooh	Cinderella	(to) go cheap
1	**dancer.n.01**	**exclaim.v.01**	**mythical person**	to overdo
2	**educator.n.01**	utter.v.02	narrative prose	to price
3	**performer.n.01**	sound.v.02	wizard, magician, sorcerer	to buy
4	**teacher.n.01**	breathe.v.01	sorceress	to overestimate
5	ballet_dancer.n.01	tremble.v.01	human being	to lower
6	attendant.n.01	murmur.v.01	**fairy tale**	to increase
7	orator.n.01	impress.v.02	nobleman	to pay
8	schoolteacher.n.01	shout.v.02	short story	**to sell**
9	chaperon.n.01	talk.v.02	female, woman	to overcharge
10	principal.n.02	**express.v.02**	poor person	**to act**
Ground truth	dancer.n.01, performer.n.01	express.v.02, exclaim.v.01	fairy tale, literary fairy tale	to sell, to deliver posession
	educator.n.01, teacher.n.01		imaginary creature, mythical person	to make a mistake, to act

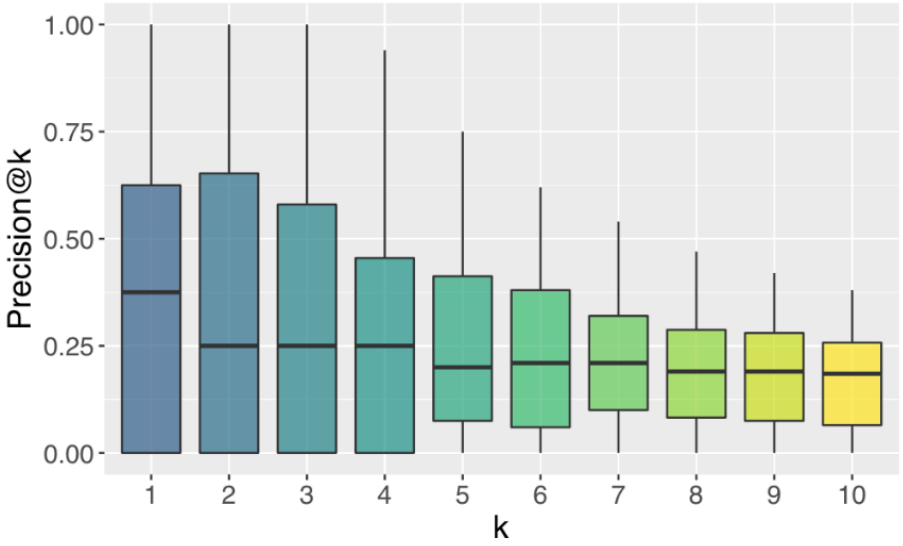
Table 4: Predicted hypernym synsets from WordNet (*dancing-master, ooh*) and RuWordNet (*Cinderella, go cheap*). Underlined green bold text denotes predictions of the model from the ground truth.

Manual evaluation result

Language	Word List
English	falanga, venerability, ambulatory, emeritus, salutatory address, eigenvalue of a matrix, liposuction, moppet, dinette, snoek, to fancify, to google, to expense, to porcelainize, to junketeer, to delist, to podcast, to deglaze, to shoetree, to headquarter
Russian	барабашка, листинг, стихосложение, аукционист, точилка, гиперреализм, серология, огрызок, фен, марикультура, уломать, отфотошопить, тяпнуть, растушевать, завраться, леветь, мозолить, загоститься, распеваться, оплавить

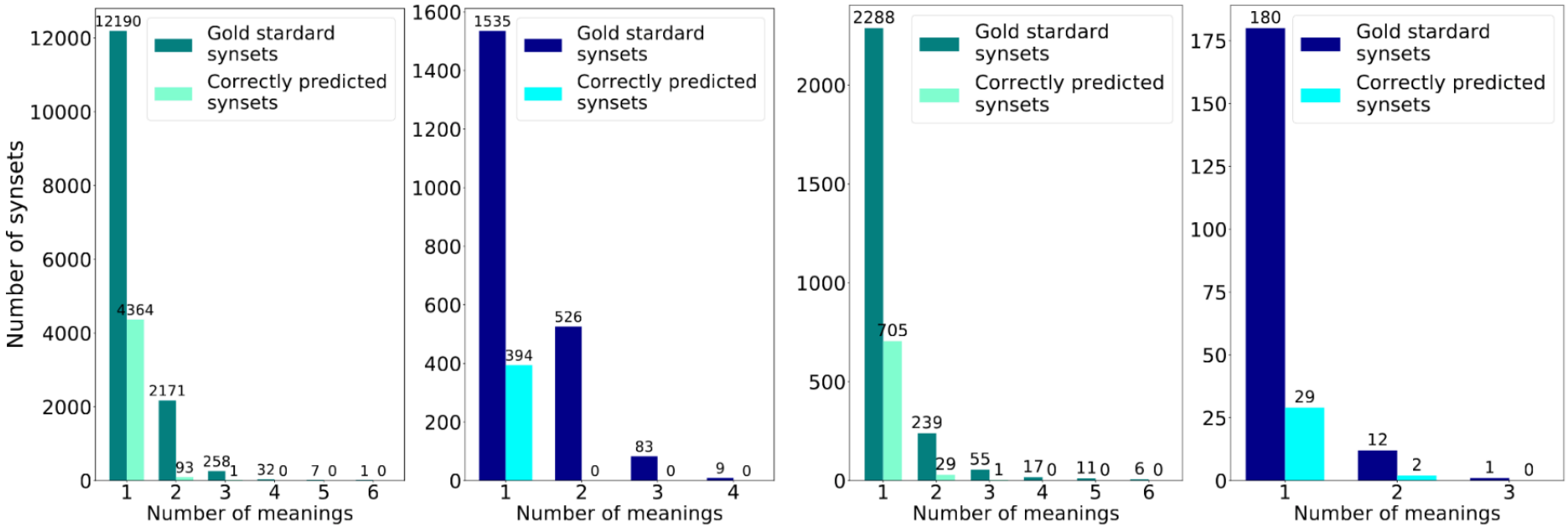


(a) Russian dataset



(b) English dataset

Why BERT did not help here: Analyzing the number of senses



(a) Russian dataset (nouns and verbs)

(b) English dataset (nouns and verbs)

Error types

Type 1. Extracted nearest neighbours can be semantically related words but not necessary co-hyponyms:

- delist (WordNet); expected senses: get rid of; predicted senses: remove, delete;
- хэштег (hashtag, RuWordNet); expected senses: отличительный знак, пометка (tag, label); predicted senses: символ, короткий текст (symbol, short text).

Type 2. Distributional models are unable to predict multiple senses for one word:

- latakia (WordNet); expected senses: tobacco; municipality city; port, geographical point; predicted senses: tobacco;
- запорожец (zaporozhets, RuWordNet); expected senses: житель города (citizen, resident); марка автомобиля, автомобиль (car brand, car); predicted senses: автомобиль, мототранспортное средство, марка автомобиля (car, motor car, car brand).

Error types

Type 3. System predicts too broad / too narrow concepts:

- midweek (WordNet); expected senses: day of the week, weekday; predicted senses: time period, week, day, season;
- медянка (smooth snake, RuWordNet); expected senses: неядовитая змея, уж (non-venomous snake, grass snake); predicted senses: змея, рептилия, животное (snake, reptile, animal).

Type 4. Incorrect word vector representation: nearest neighbours are not semantically close:

- falanga (WordNet); expected senses: persecution, torture; predicted senses: fish, bean, tree, wood.;
- кубокилометр (cubic kilometer, RuWordNet); expected senses: единица объема, единица измерения (unit of capacity, unit of measurement); predicted senses: город, городское поселение, кубковое соревнование, спортивное соревнование (city, settlement, competition, sports contest).

Type 5. Unaccounted senses in the gold standard datasets, inaccuracies in the manual annotation:

- emeritus (WordNet); expected senses: retiree, non-worker; predicted senses: professor, academician;
- сепия (sepia, RuWordNet); expected senses: морской моллюск "sea mollusc"; predicted senses: цвет, краситель (color, dye).

Conclusion

- Taxonomy Enrichment task:
  - Add new words to existing lexical-semantic resource RuWordNet and WordNet.
  - Diachronic

- The participants used various types of information
  - Embeddings (word2vec, fasttext, BERT)
  - Electronic lexicons
  - Search engines representation pages
- A prominent direction for future work
  - Integration of graph embeddings of the taxonomy with the distributional information
  - Application for construction of taxonomies

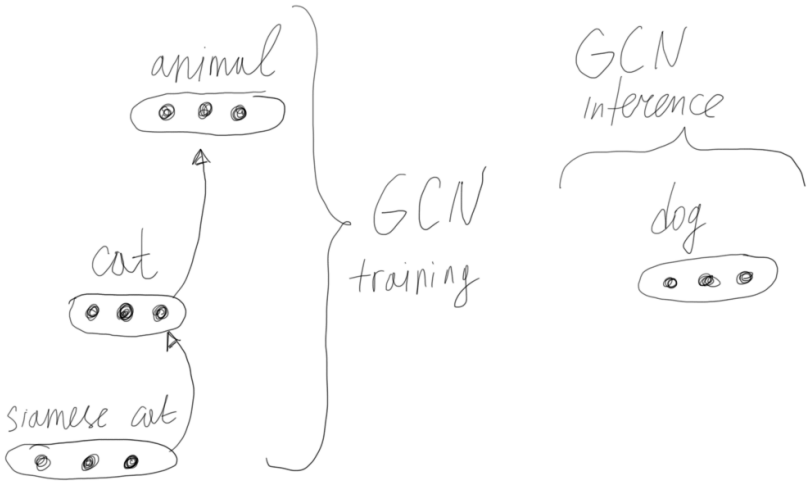
Graph representations for taxonomy enrichment

Nikishina, I., Logacheva, V., Panchenko, A., and Loukashevich N. (2021): Exploring Graph-based Representations for Taxonomy Enrichment. Global WordNet Conference. Pretoria. South Africa

- Link Prediction Using GCN Autoencoder:
  1. Use the graph autoencoder model (Kipf and Welling, 2016).
  2. FastText embeddings as features, relations for ego network.
  3. Get its vector representation from the encoder.
  4. Predict the probability of the link between the new node and all other nodes in the graph.
  5. The top-10 synsets from the existing taxonomy are used.

method	nouns			verbs		
	1.6-3.0	1.7-3.0	2.0-3.0	1.6-3.0	1.7-3.0	2.0-3.0
Poincaré embeddings	0.0593	0.0658	0.1013	0.1255	0.0656	0.1092
node2vec (top-5 fastText associates)	0.1938	0.2187	0.1554	0.1514	0.1091	0.1469
node2vec (projection)	0.0400	0.0273	0.0218	0.1041	0.0517	0.0377
GCN autoencoder	0.1570	0.1751	0.1677	0.1088	0.0937	0.1173

method	nouns		verbs	
	non-restricted	restricted	non-restricted	restricted
Poincaré embeddings	0.1431	0.2517	0.1050	0.1397
node2vec (top-5 fastText associates)	0.2660	0.3659	0.1681	0.2518
node2vec (projection)	0.1854	0.2527	0.1800	0.2531
GCN autoencoder	0.1826	0.2605	0.0948	0.1406



Graph representations for taxonomy enrichment

Shen et al. (2020): TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network https://arxiv.org/pdf/2001.09522.pdf