

Computational Linguistics and Intellectual Technologies:  
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

## COMBINED APPROACH TO HYPERNYM DETECTION FOR THESAURUS ENRICHMENT

**Tikhomirov M. M.** (tikhomirov.mm@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

**Loukachevitch N. V.** (louk\_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia,  
Kazan Federal University, Kazan, Russia

**Parkhomenko E. A.** (parkat13@yandex.ru)

Lomonosov Moscow State University, Moscow, Russia

This paper describes a combined approach to hypernym detection task. The approach combines the following techniques: distribution semantics, rule-based patterns, and modern neural networks (BERT). An important feature of our solution is that hypernyms are extracted only from a single text collection provided by the organizers. The described approach obtained the fourth result on the private nouns track. It was found out that the use of the rule-based patterns can significantly improve the results. Also, using the BERT model as an additional factor always helps to improve the performance.

**Key words:** hypernym extraction, thesaurus, embeddings, patterns, BERT

**DOI:** 10.28995/2075-7182-2020-19-736-746

# КОМБИНИРОВАННЫЙ ПОДХОД К ИЗВЛЕЧЕНИЮ ГИПЕРОНИМОВ ДЛЯ ОБОГАЩЕНИЯ ТЕЗАУРУСА

**Тихомиров М. М.** (tikhomirov.mm@gmail.com)

МГУ имени М. В. Ломоносова, Москва, Россия

**Лукашевич Н. В.** (louk\_nat@mail.ru)

МГУ имени М. В. Ломоносова, Москва, Россия,  
Казанский Федеральный Университет, Казань, Россия

**Пархоменко Е. А.** (parkat13@yandex.ru)

МГУ имени М. В. Ломоносова, Москва, Россия

В данной статье описывается комбинированный подход к решению задачи извлечения гиперонимов. В работе используются: методы дистрибутивной семантики, шаблоны на основе правил и современные нейронные сети (BERT). Важной особенностью описанного решения является то, что извлечение гиперонимов происходит только на основе коллекции текстов, предоставленной организаторами. Описанный подход получил четвертый результат в задаче предсказания гиперонимов существительных. Было показано, что использование шаблонов на основе правил может значительно улучшить результаты. Кроме того, использование BERT в качестве дополнительного фактора всегда улучшает результаты.

**Ключевые слова:** извлечение гиперонимов, тезаурус, векторные представления, шаблоны, BERT

## 1. Introduction

Knowing relationships between words such as a hypernym (more general word) or hyponym (more specific word) can be useful in various tasks of natural language processing such as question answering, sentiment analysis, textual entailment, etc. Lexical relationships can be found in lexical-semantic resources such as WordNet [13], BabelNet [14] and others. However, development and maintenance of large semantic resources is a very difficult, time-consuming activity. Besides, such resources are never comprehensive: some relations significant for a specific domain or a task can be absent in the used resource. Numerous studies have been devoted to the problem of hypernym detection from text collections or definitions in conventional vocabularies [7], [22], [6], [16]. The researchers utilized such methods as the use of linguistic patterns, unsupervised distributional approaches, machine learning, and combinations of several methods.

For evaluation, the task of hypernym detection can be formulated as a two-class classification of relations. Some approaches attempted to evaluate the whole structure of the taxonomy created on the basis of extracted hypernyms [3], [19]. The organizers of the SemEval 2018 hypernym detection task [4] considered the task of hypernym detection as a ranking problem when correct hypernyms had to be located closer to the top of the hypernym candidate list.

In 2020, the first evaluation for extracting hypernyms for Russian RUSSE-2020 Taxonomy Enrichment was organized [15]. The task was to provide the most appropriate hypernym synsets from the published version of RuWordNet [11] for new words. Previously, a new, extended version of RuWordNet had been prepared manually but not published. It was proposed to use extended RuWordNet as a gold standard in evaluation of hypernym extraction.

The organizers provided a news collection to participants ensuring that the practically all novel words under consideration occurred in the collection with frequency more than 50, but no restrictions were placed on the use of additional resources. Solving such a task, the participants could consider two possible approaches in utilizing data. The first approach is to use various available resources (text collections, vocabulary definitions, pre-trained embeddings) to detect hypernym synsets for a given word. The second approach can be to use mainly the provided corpus trying to extract all possible relations from it. In the current paper, we consider the second approach: in our experiments we used only the given text collection. We supposed the following practical scenario: for tuning a general resource to a specific domain, only a specialized text collection can be given, other types of available data can be too general or irrelevant to the domain.

We consider a combined approach to hypernym extraction from a given text collection including distributional representation of words and phrases, patterns, accounting named entities, which have to be distinguished from ordinary words, and the BERT model [5] trained for hypernym prediction.

## 2. Related Work

Approaches for hypernym extraction from texts can be subdivided into: pattern-based (linguistic) approaches, unsupervised distributional approaches, machine learning methods, and combined approaches.

Linguistic approaches are based on obtaining relations between words revealing their joint occurrences in a certain set of patterns [7]. The use of these methods presupposes that if some specified context is detected between entities  $X$  and  $Y$ , then, there is a given relationship between these entities. For the taxonomic relationship, such a context can be the context “ $X$  is a kind of  $Y$ ”. The pattern-based approach is highly accurate and simple, and can be applied to various types of relationships. Later pattern approaches exploited syntactic structures [21], [18]. In [20], the authors trained the LSTM neural network to generate embeddings for syntactic paths for hypernym extraction. In Russian, Sabirova and Lukanin [17] described patterns for hypernym extraction from Russian texts.

Distributional approaches are based on vector representations of words. Several methods were used to represent term relationships as a combination of each

term’s embedding vector: concatenation [1], difference [16], and dot-product. The best results using distributional methods were achieved with supervised approaches. Supervised projection learning models are trained to project the embedding of the target word such that its projection is close to that of its hypernym [6], [23]. Fu et al. [6] proposed also to learn multiple projection matrices representing different kinds of hypernymy relationships. Using this approach, Yamane et al. [25] presented a model that jointly learns word clusters and corresponding projection matrices.

However, Levy et al. [9] revealed that the results achieved by supervised distributional models are mainly due to so-called “lexical memorization phenomenon”, that is the models learn that  $y$  is a prototypical hypernym regardless of  $x$ . For example, if a classifier obtains many positive examples with the word  $y = \text{animal}$ , it may learn that anything that appears with  $y = \text{animal}$  should generate the positive answer. Based on these results, Camacho-Collados et al. [4] proposed to change the evaluation scheme of hypernym extraction from two-class classification to the ranking scheme.

The advantages of different methods are combined in incremental taxonomy generation [22], [26], in which all of the above characteristics (occurrence in templates, similarities in different types of contexts (linear and syntactic, local and global), joint occurrence) serve as features for a classifier that decides to join the next word to the created taxonomy.

The CRIM system, the winner of the SemEval 2018 Hypernym Discovery task, exploited the combined approach to hypernym recognition [2]. Their approach comprised hypernym patterns, co-hyponym patterns, and supervised projection learning. Held and Habash [8] practically repeated and partially improved the results of the CRIM system using a significantly simpler combined method based on hypernym patterns and distributional similarity of a term from the test collection to the most similar term from the training collection.

### 3. Data and Task

The task of RUSSE-2020 taxonomy enrichment is to find the closest synset from the published RuWordNet version for a novel word. The gold standard data are taken from prepared but not currently published extended RuWordNet version. Only the closest hypernym from RuWordNet for a given word or its direct hypernyms are considered as correct answers.

The task comprises nouns and verbs. There were 762 nouns and 175 verbs in the public set for developing and tuning algorithms, and 1525 nouns and 350 verbs in the private (test) data. The evaluation was organized in the CodaLab evaluation system<sup>1</sup>. Besides, the organizers provided the training dataset constructed on published RuWordNet, a news collection of 2017 (further, News2017 corpus) gathered from more than 1000 news sources, in which new words under evaluation mainly occurred at least 50 times, and some other resources.

This news text collection consists of 8 million articles, a total of 2.2 billion tokens. Statistics of the target words can be seen in **Table 1**. It can be seen that the

---

<sup>1</sup> <https://competitions.codalab.org/competitions/22168>

corpus provides more than 50 occurrence for almost all words in the datasets. Therefore it is possible to attempt to obtain hypernyms for words in the current task from this specific corpus without using additional sources of candidates.

**Table 1:** Target words statistics

	median	max	freq < 50
public nouns	188.5	133,683	10
public verbs	128.5	16,657	7
private nouns	183.0	140,173	28
private verbs	140.5	45,807	14

Participating system have to generate a list of the most relevant hypernym synsets for novel words. The top 10 synsets are considered as answers. The main measure is mean average precision MAP, which is equal to 1 as maximum when all the correct answers are at the top of the candidate list. The second used measure is MRR, which is calculated on the basis of the first correct answer.

## 4. Method for Hypernym Extraction

Our method of hypernym extraction is intended to extract maximum information from a given text collection, in the current case from the News2017 text collection provided by the organizers. The method comprises the following components:

- distributional (vector) representation of words and phrases under analysis,
- linguistic patterns for hypernyms and co-hyponyms,
- special processing of named entities to remove their contexts from consideration,
- application of BERT in a supervised manner to approve hypernym candidates.

### 4.1. Word embeddings

In this study word embeddings were trained only on the News2017 corpus. As a training method we used not word2vec, but the traditional approach in distributional semantics, which includes the following steps:

1. The matrix of co-occurrence frequencies of words in the corpus is calculated,
2. Scores of word co-occurrences are recalculated using the positive point-wise mutual information (PPMI) measure [10],
3. SVD method over PPMI matrix is applied, which allows reducing the dimension of the matrix from the vocabulary size to a chosen smaller value,
4. When processing the corpus, phrases from the thesaurus were merged into united tokens, for example **rice\_grain**. The models were calculated with 600 vector size and different window sizes: 1, 3, 5.

The PPMI + SVD approach demonstrated comparable quality with word2vec in a number of experiments [10] and showed better results in our internal studies.

For all target words, top 100 word candidates from each model were considered, based on cosine similarity. Also, words must be present in the thesaurus. Then, the list of candidate synsets was formed by extracting hypernyms for synsets and hypernyms of hypernyms (with penalty weight) corresponding to candidate words. For each candidate synset, the average cosine similarity of the initial words (*cos\_sim\_list*) and the number of times the algorithm returned this synset (*count*) were calculated; the candidates were ranked by the following formula:

$$base_{score} = mean(cos\_sim\_list) \cdot \log_2(1 + count) \cdot \alpha, \quad (1)$$

where  $\alpha$  is equal to 0.3, if estimated synset is hypernym of hypernym.

## 4.2. Excluding Named Entity Contexts

It was found in previous works that named entities can include general words and in such a way to distort the contexts of ordinary words [2], [12]. For example, Loukachevitch [12] found that Russian word *mistral*, which means “a strong, cold, north-westerly wind. in the northern Mediterranean” and is linked to hypernym *wind* in Ru-WordNet, but in current news articles in Russian, this word mainly means the class of French helicopter carriers. We assumed that the RUSSE evaluation dataset contains mainly ordinary words and decided to train word embeddings excluding named entity contexts from consideration. Examples of entries in the RUSSE dataset that coincide with named entities are as follows: *trepak* (dance vs. family name), *chub* (strand of hair vs. family name), *ryabinovka* (alcoholic beverage vs. names of villages), and others.

Currently, we applied the following simple preprocessing procedure: if a word occurrence is capitalized and it is not the first word in a sentence then such an occurrence obtains a special prefix, which means that the occurrence and its contexts are excluded from calculating embeddings.

## 4.3. Use of Patterns

In addition to the basic method, two kinds of patterns were extracted and used. All patterns were applied to the pairs of the target word and all similar words extracted by a distributional model. The appearance of candidates in the patterns increases the similarity weight of the candidate words to the target word, because pattern matching is an additional evidence of semantic similarity. For example, it is examined whether pattern matching exist for the target word *peony* and the candidate word *rose*. If it exists, then it is evidence, that *rose* and *peony* can be synonyms or co-hyponyms.

All patterns were automatically extracted using regular expressions on the News2017 corpus. Two types of patterns were considered:

1. co-hyponym patterns, whose successful matching leads to an increase in the weight of hypernym synsets for words from top most similar words,
2. synonym-hypernym patterns, whose successful matching leads to both: an increase in the weight of hypernym synset and an additional inclusion in the candidate list of a direct synset of a candidate word, but not just its hypernym as in the base model.

The following are sample patterns (splitted by ; character), where X is a target word, Y is a candidate word, and W is any word.

- Examples of co-hyponym patterns: X , W , Y; Y, W, X; Y , X; X, Y; Y and X; X and Y; Y or X; X or Y;
- Examples of synonym—hypernym patterns: X ( Y; Y ( X; Y—X; X—Y; X—is Y; Y—is X; X , W and another Y; X and W—is type of Y.

Examples of co-hyponyms extracted with the described patterns for word *ice-cream* (with frequencies in the corpus) are as follows: *chocolate* (145), *sweets* (106), *candy* (93), *cookie* (69), *yogurt* (61), .. Examples of pattern-based hypernyms for *ice-cream* include the following words: *delicacy* (70), *dessert* (21), *cake* (13), *sweets* (10), ...

When applying a co-hyponym pattern, the frequency of being in one sentence for target and candidate words, is calculated. The modified formula is as follows:

$$upd\_pattern\_score = base\_score \cdot (1 + shp\_hit) \cdot (1 + \frac{2 \cdot chp\_count}{os\_count + 2}) \quad (2)$$

where *shp\_hit* is equal to 1 if the synonym-hypernym pattern was matched, *chp\_count* is a count of times when co-hyponym pattern was matched, and *os\_count* is a count when words were in one sentence.

In addition to modifying the formula, direct synsets (synsets of the most similar words) were also added as candidates, when the synonym-hypernym pattern was matched.

#### 4.4. Use of BERT to Assess Hypernym Candidate

A neural network architecture such as BERT [5] can be used for the hypernym prediction task [27]. BERT is a transformer encoder [26], and one of its key features is that it is trained on a large amount of unlabeled data. After this procedure, the model shows strong results on a wide range of specific tasks.

For the hypernym prediction task, the following approach was implemented:

1. The binary classification problem was considered;
2. The BERT input was a pair of words with special characters: [CLS] word1 [SEP] word2 [SEP]; word2 can be a multi-word expression. If word2 is a hypernym of word1, then the label is 1, and 0 otherwise;
3. The training data was created from RuWordNet. For each positive example, three negative examples were added. The negative examples were uniformly sampled from: random synsets, hypernyms of hypernyms, hyponyms and hyponyms of hypernyms of word1.

The model was fine-tuned on 5 epochs with 2e-4 learning rate for the classification layer and 2e-5 for the BERT layers. The training dataset consisted of 1.5 million examples.

Further, for each candidate synset, its probability of being a hypernym and the maximum probability of its hyponyms to be a hypernym to the target word were calculated. When calculating the probability of synset, the probabilities of all synset entries were calculated, and then averaged. The resulting BERT probability is:

$$bert\_prob = 0.6 \cdot syn\_bert\_prob + 0.4 \cdot max\_hyp\_syn\_bert\_prob \quad (3)$$

And then, the final modified formula for ranking is as follows:

$$\text{synset\_score} = \text{upd\_pattern\_score} \cdot (1 + \text{bert\_prob}) \quad (4)$$

For example, for the word *agnostic* there is a hypernym synset with name *follower*. In this case, BERT probability will be 0.99, model predicts such a probability directly for a given synset.

## 5. Results

In this competition, for each target word, it was necessary to provide a list of 10 candidate hypernym synsets, ranked from more probable to less probable. To assess the quality of the models, the mean average precision (MAP) and mean reciprocal rank (MRR) metrics were used.

**Table 2:** Nouns track results

	public		private	
	MAP	MRR	MAP	MRR
base	0.446	0.478	0.440	0.474
base-ne	0.449	0.482	0.444	0.478
base-ne, co-hyponym patterns	0.473	0.509	0.467	0.502
base-ne, hypernym patterns	0.467	0.505	0.456	0.494
base-ne, both patterns	0.482	0.522	0.481	0.520
base-ne, bert	0.471	0.507	0.455	0.490
base-ne, bert, co-hyponym patterns	0.494	0.533	0.475	0.509
base-ne, bert, hypernym patterns	0.484	0.524	0.470	0.508
base-ne, bert, both patterns	0.509	0.550	0.493	0.531

**Table 3:** Verbs track results

	public		private	
	MAP	MRR	MAP	MRR
base	0.254	0.295	0.255	0.291
base, co-hyponym patterns	0.285	0.327	0.246	0.286
base, hypernym patterns	0.260	0.303	0.254	0.290
base, both patterns	0.288	0.332	0.244	0.284
base, bert	0.277	0.322	0.265	0.305
base, bert, co-hyponym patterns	0.314	0.358	0.259	0.302
base, bert, hypernym patterns	0.288	0.334	0.264	0.304
base, bert, both patterns	0.324	0.371	0.254	0.296

For verbs and nouns, slightly different parameters were applied. For nouns, embeddings with only window 1 were considered; for verbs, embeddings with windows 1, 3, and 5 were combined. In case when named entity contexts were excluded, the

results of the models with and without this preprocessing were also combined by averaging of similarities. The configuration called **base** refers to the basic approach described in [section 4.1](#), and the **base-ne** configuration refers to **base** configuration with names excluding, described in [section 4.2](#). [Table 2](#) describes the results of the proposed methods for nouns, and [Table 3](#) contains the results for verbs.

The following conclusions can be drawn from the results:

- Exclusion of named entity contexts slightly improves the results;
- In all cases, except for private verbs, matching patterns significantly improves the results. The problem with private verbs will be investigated later;
- Using BERT in the described way always improves the results.

The described approach obtained the fourth result on the noun private set without any additional vocabulary definitions, used by the first two approaches (according to the provided descriptions) and is very close to the results of the third place approach. In practice, in specific domains external vocabularies can be irrelevant to the domain and provided text collection. The use of the combined method improved the hypernym predictions for nouns by more than 10% in the public and private sets if compared to the basic model.

For verbs, the improvement on the public set was also significant (more than 25%) but in the private dataset the improvement was not reproduced.

## 6. Conclusion

The paper describes a combined approach to the hypernym detection task. The approach combines the following techniques: distributional semantics, rule-based patterns, and modern neural. Hypernyms are extracted only from the text collection provided by the organizers. This reflects the real situation when it is necessary to expand the existing thesaurus to a new specific domain.

The described approach obtained the fourth result on the private nouns track, with 0.493 MAP and 0.531 MRR. It was found that the use of the described patterns can significantly improve the results. Also, using the BERT model as an additional factor always helps to improve the results. The exclusion of name contexts for target words also helps to improve the results of the system, by improving the results on specific words.

As further research, it is necessary to more deeply explore the possibilities of using BERT for this task. In addition, it is planned to investigate the unstable behavior of patterns on verbs. Also, the behavior of the system strongly depends on the algorithm for obtaining candidates, so it is planned to explore various ways of doing this.

## 7. Acknowledgements

The research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University. The participation of M. Tikhomirov in the reported study was funded by RFBR, project number 19-37-90119. The work of N. Loukachevitch concerning formulation of main principles of work with

a text collection, accounting named entities, and patterns is funded by Russian Science Foundation, research project № 19-71-10056, financed through Kazan Federal University.

## References

1. *Baroni M., Lenci A.* (2011), How we BLESSED distributional semantic evaluation, Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, Association for Computational Linguistics, Edinburgh, pp. 1–10.
2. *Bernier-Colborne G., Barriere C.* (2018), CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery, Proceedings of The 12th International Workshop on Semantic Evaluation, New Orleans, pp. 725–731.
3. *Bordea G., Lefever E., Buitelaar P.* (2016) Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, pp. 1081–1091.
4. *Camacho-Collados J., Delli Bovi C., Espinosa-Anke L., Oramas S., Pasini T., Santus E., ... & Saggion H.* (2018). SemEval-2018 task 9: Hypernym discovery, Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, pp. 712–724.
5. *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.* (2018), Bert: Pre-training of deep bidirectional transformers for language understanding, available at <https://arxiv.org/abs/1810.04805>.
6. *Fu, R., Guo, J., Qin, B., Che, W., Wang, H., & Liu, T.* (2014), Learning semantic hierarchies via word embeddings, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1199–1209.
7. *Hearst M.* (1992), Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the 14th conference on Computational linguistics-Volume 2, pp. 539–545.
8. *Held W., Habash N.* (2019), The Effectiveness of Simple Hybrid Systems for Hypernym Discovery, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3362–3367.
9. *Levy, O., Remus, S., Biemann, C., & Dagan, I.* (2015), Do supervised distributional methods really learn lexical inference relations?, Proceedings of NAACL-HLT 2015, pp. 970–976.
10. *Levy, O., Goldberg, Y., Dagan, I.* (2015), Improving distributional similarity with lessons learned from word embeddings, Transactions of the Association for Computational Linguistics, Vol. 3, pp. 211–225.
11. *Loukachevitch, N. V., Dobrov, B. V.* (2014), RuThes linguistic ontology vs. Russian wordnets, Proceedings of the Seventh Global Wordnet Conference, pp. 154–162.
12. *Loukachevitch N. V.* (2019), Corpus-based Check-up for Thesaurus, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5773–5779.
13. *Miller, G.* (1998), Nouns in WordNet, WordNet – An Electronic Lexical Database, Fellbaum, C (ed), The MIT Press, pp. 23–47

14. Navigli R., Ponzetto S. P. (2010), BabelNet: Building a very large multilingual semantic network, Proceedings of the 48th annual meeting of the association for computational linguistics, pp. 216–225.
15. Nikishina I., Logacheva V., Panchenko A., Loukachevitch N. (2020), RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian language, Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2020.
16. Roller S., Erk K., Boleda G. (2014), Inclusive yet selective: Supervised distributional hypernymy detection, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1025–1036.
17. Sabirova K., Lukanin A. (2014), Automatic Extraction of Hypernyms and Hyponyms from Russian Texts, Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST 2014), pp. 35–40.
18. Sang E. T. K., Hofmann K. (2009), Lexical patterns or dependency patterns: which is better for hypernym extraction? ,Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pp. 174–182.
19. Sarkar, R., McCrae, J.P., Buitelaar, P. (2018), A supervised approach to taxonomy extraction using word embeddings, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)
20. Shwartz V., Dagan I. (2016), Path-based vs. distributional information in recognizing lexical semantic relations, COLING, pp. 24–29.
21. Snow R., Jurafsky D., Ng A. Y. (2004), Learning syntactic patterns for automatic hypernym discovery, Proceedings of Advanced in Neural Information Processing systems (NIPS), pp. 1297–1304.
22. Snow R., Jurafsky D., Ng A. Y. (2006), Semantic taxonomy induction from heterogeneous evidence, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 801–808.
23. Ustalov, D., Arefyev, N., Biemann, C., & Panchenko, A. (2017), Negative Sampling Improves Hypernymy Extraction Based on Projection Learning, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, pp. 543–550.
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), Attention is all you need, Advances in neural information processing systems, pp. 5998–6008.
25. Yamane, J., Takatani, T., Yamada, H., Miwa, M., & Sasaki, Y. (2016), Distributional hypernym generation by jointly learning clusters and projections, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1871–1879.
26. Yang H., Callan J. (2009), A metric-based framework for automatic taxonomy induction, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AF-NLP: Volume 1-Volume 1, Association for Computational Linguistics, pp. 271–279.
27. Yao L., Mao C., Luo Y. (2019), KG-BERT: BERT for Knowledge Graph Completion, available at <https://arxiv.org/pdf/1909.03193.pdf>.