Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection

Maxim Rachinskiy
HSE University
Moscow, Russia
myurachinskiy@edu.hse.ru

Nikolay Arefyev
Samsung Research Center Russia
Lomonosov Moscow State University
HSE University
Moscow, Russia
narefjev@cs.msu.ru

Abstract

Consulting word definitions from a dictionary is a familiar way for a human to find out which senses a particular word has. We hypothesize that a system that can select a proper definition for a particular word occurrence can also naturally solve Semantic Change Detection (SCD) task. To verify our hypothesis, we followed an approach previously proposed for Word Sense Disambiguation (WSD) and trained a system that embeds word definitions and word occurrences into the same vector space. In this space, the embedding of the most appropriate definition has the largest dot product with a contextualized word embedding.

The system is trained on an English WSD corpus. To make it work for the Russian language, we replaced BERT with the multilingual XLM-R language model and exploited its zero-shot cross-lingual transferability. Despite not finetuning the encoder model on any Russian data, this system achieves the second place in the competition, and likely works for any of one hundred other languages XLM-R was pre-trained on, though the performance may vary. We then measure the impact of such WSD pre-training and show that this procedure is crucial for our results. Since our model was trained to choose a proper definition for a word, we propose an algorithm for the interpretation and visualization of the semantic changes through time.

By employing additional labeled data in Russian and training a simple regression model, that converts the distances between output contextualized embeddings into more human-like scores of sense similarity between word occurrences, we further improve our results and achieve the first place in the competition.

Keywords: Semantic change detection, SCD, gloss-informed models, GLM
DOI: 10.28995/2075-7182-2021-20-578-586

Межъязыковой перенос без дообучения толковой языковой модели для обнаружения семантических сдвигов

Рачинский Максим[◇]
Москва, Россия
myurachinskiy@edu.hse.ru

Арефьев Николай[△▽◇]
Москва, Россия
narefjev@cs.msu.ru

[◇]Национальный исследовательский университет «Высшая школа экономики»
[△]Московский Исследовательский Центр Самсунг
[▽]Московский Государственный Университет им. М. В. Ломоносова

Аннотация

Обращение к определениям из словаря — это привычный для человека способ выяснить, какие значения имеет то или иное слово. Мы предполагаем, что система, которая может выбрать из толкового словаря или глоссария правильное определение для конкретного вхождения слова, также может естественным образом решить задачу обнаружения изменений значений слов с течением времени (семантических сдвигов). Чтобы проверить нашу гипотезу, мы использовали подход, ранее предложенный для разрешения лексической многозначности (WSD), и обучили систему, которая проецирует определения слов и их вхождения в тексты в одно и то же векторное пространство. В этом пространстве вектор наиболее подходящего определения имеет самое большое скалярное произведение с контекстуализированным вектором вхождения слова.

Система обучается разрешать лексическую многозначность (выбирать самое подходящее определение) на англоязычном корпусе. Для того чтобы работать с текстами на русском языке, мы заменили англоязычный BERT на многоязычную языковую модель XLM-R и использовали ее способность к межъязыковому переносу. Несмотря на отсутствие дообучения модели на каких-либо данных на русском языке, такая система заняла второе место в соревновании и, вероятно, работает на любом из ста других языков, на которых

XLM-R был предварительно обучен, хотя в зависимости от языка качество может варьироваться. Мы оцениваем влияние обучения модели выбору наиболее подходящего определения и показываем, что эта процедура имеет решающее значение для наших результатов. Поскольку наша модель была обучена подбору правильного определения слова, мы используем это свойство и предлагаем метод интерпретации и визуализации семантических сдвигов во времени.

Используя дополнительные размеченные данные на русском языке и обучая простую регрессионную модель, которая преобразует расстояния между контекстуализированными векторами вхождений слов в оценки смыслового сходства, близкие к человеческим, мы улучшили наши результаты. Дообученная на русскоязычных данных система заняла первое место в соревновании.

Ключевые слова: семантические сдвиги, модели на основе определений

1 Introduction

RuShiftEval [6] is a semantic change detection task for the Russian language.[1] Each test sample in the competition consisted of a single Russian word. The participants were asked to predict how much test words have changed their meanings between three epochs: pre-Soviet, Soviet and post-Soviet. The mean of the three Spearman correlation coefficients of the predicted and gold scores was utilized as the main performance metric.

Through the evaluation period, our model which did not use any Russian data for finetuning achieved the second place in the competition. As there was no domain adaptation, this system likely works for any of one hundred other languages XLM-R was pre-trained on, though the performance may vary. After adding labeled data in Russian and training a simple regression model, that converts the distances between output contextualized embeddings we achieved the first place.[2]

Our main interest was whether the semantic change detection systems can benefit from using gloss information and how these systems can be interpreted.

2 Background

Here we summarize the prior work linking word occurrences and word definitions. One of the first approaches in this field [7] calculated the lexical overlap between the context of a particular word occurrence and all possible definitions of this word. This approach did not take into account word synonymy or other lexical relations. The recent works tried to combine state-of-the-art language models with glosses from some dictionaries.

One of such methods has been proposed in [15]. Their EWISE system used a pre-training procedure for a gloss encoder, that learned knowledge graph embeddings from WordNet [8]. After this pre-training, the authors froze the gloss encoder and started to train a context encoder with labeled WSD data. The ablation study of this work has shown the importance of such gloss encoder pre-training.

While the previous method requires relational information from a knowledge graph, the method proposed in [5] relies fully on gloss information. The developed system jointly encodes the context with all possible glosses of the target word. The authors used a pre-trained BERT [1] model as initialization for their encoder. The results demonstrated a big gap in the performance between the developed method and a simple context encoder without any gloss information.

A similar approach has been proposed in [2], where authors trained two separate Transformer-based encoders for word occurrences (Context encoder) and word definitions (Gloss encoder), both initialized with BERT weights [1]. To represent a word occurrence, the outputs of the Context encoder for all of its subwords were averaged. To represent a definition, the output of the Gloss encoder from [CLS] token was taken. Finally, for a word occurrence and all of its definitions, the dot products between those outputs were calculated and the softmax function was applied to them, resulting in a probability distribution over possible word senses. The whole model was trained using cross-entropy loss to select the correct word sense on WSD data.

While BEM [2] and GlossBERT [5] are based on BERT [1] encoder, our system exploits XLM-RoBERTa (XLM-R) [13] architecture. XLM-R is based on RoBERTa [10] and is pre-trained on unlabeled

[1] `https://competitions.codalab.org/competitions/28340`

[2] In order to make our results reproducible, we publish the code of our experiments: `https://github.com/myrachins/RuShiftEval`.

ID	Model	P1	P2	P3	Aver.
GLM, zero-shot cross-lingual transfer to Russian					
1	Manhattan+norm GLM xlmr.large	74.1	77.9	79.8	77.3
GLM + regression to human scores trained on RuSemShift					
2	Linear regression on GLM xlmr.large distances	77.0	80.1	81.8	79.6
3	Linear regression on GLM xlmr.large+base distances	78.1	**80.3**	**82.2**	**80.2**
4	Knn regression on GLM xlmr.large+base distances	71.8	76.2	80.9	76.3
5	Random.forest regr. (1K est.) on GLM xlmr.large+base dist.	75.2	78.7	81.6	78.5
6	Random.forest regr. (2K est.) on GLM xlmr.large+base dist.	75.0	78.7	81.7	78.5
7	Random.forest regr. (5K est.) on GLM xlmr.large+base dist.	75.8	78.4	81.3	78.5
The top3 best results of other teams					
-	DeepMistake	**79.8**	77.3	80.3	79.1
-	vanyatko	67.8	74.6	73.7	72.0
-	aryzhova	46.9	45.0	45.3	45.7

Table 1: Test Spearman score for each of our submissions. P1, P2, P3 columns stand for the pre-Soviet - Soviet, Soviet - post-Soviet and pre-Soviet - post-Soviet pairs respectively. The features for 2-7 submissions were taken from 7 different distance measures: L1, L1+norm, L2, L2+norm, Dot product, Dot product+norm(Manh), Cosine. Top-head models from 2-7 submissions were trained with RuSemShift data [11]. GLM pre-trained XLM-RoBERTa encoders were used in a feature-based setting and were not fine-tuned.

data with MLM (Masked Language Modeling) objective. But in contrast to BERT and RoBERTa, it is pre-trained not on monolingual data but on 2.5 TB of texts from CommonCrawl in 100 languages.

3 System overview

Our approach employs contextualized embeddings obtained from a gloss-based Word Sense Disambiguation (WSD) model to measure an average sense similarity between occurrences of a particular word in two corpora. This model was pre-trained with Gloss Language Modeling (GLM) procedure, which we will discuss further in detail and compare with pure Masked Language Modeling (MLM) pre-training.

Based on the previous observations, that the strongest signal in the contextualized embeddings of a language model pre-trained with MLM corresponds to the word form, not the word meaning [4], we try to fix it by fine-tuning the model to select a gloss (a definition) from WordNet, that is most appropriate for a particular word occurrence. We call this model a Gloss Language Model (GLM) and show that this training procedure results in much more appropriate contextualized embeddings for the SCD task. To initialize the GLM before training, we use the weights of the XLM-R model, which was pre-trained with MLM objective on 2.5TB of texts from 100 languages [13]. Despite using only English WSD data for training, our model still produces sensible contextualized embeddings for Russian, which alone gives a strong performance in the SCD task. Since we do not use any Russian data or resources for the pure GLM-based SCD model, this model will likely work for other languages too, though the performance may vary.

To solve the SCD task, we sample sentences containing a particular target word from each of the given three epochs. Finally, for each pair of epochs, we calculate the average distance between contextualized word embeddings from GLM.

3.1 Gloss-informed embeddings

In order to learn sense-dependent representations of words, we pre-train our system on the Word Sense Disambiguation task. Following the BEM model [2], our system consists of two separate encoders: Context Encoder and Gloss Encoder.

Context encoder (T_c) takes a sentence $c = c_0, \ldots, c_{i-1}, w_c, c_{i+1}, \ldots, c_n$ containing a target word w_c to be disambiguated, where w_c is the i^{th} word in the sentence. The encoder then produces the target word

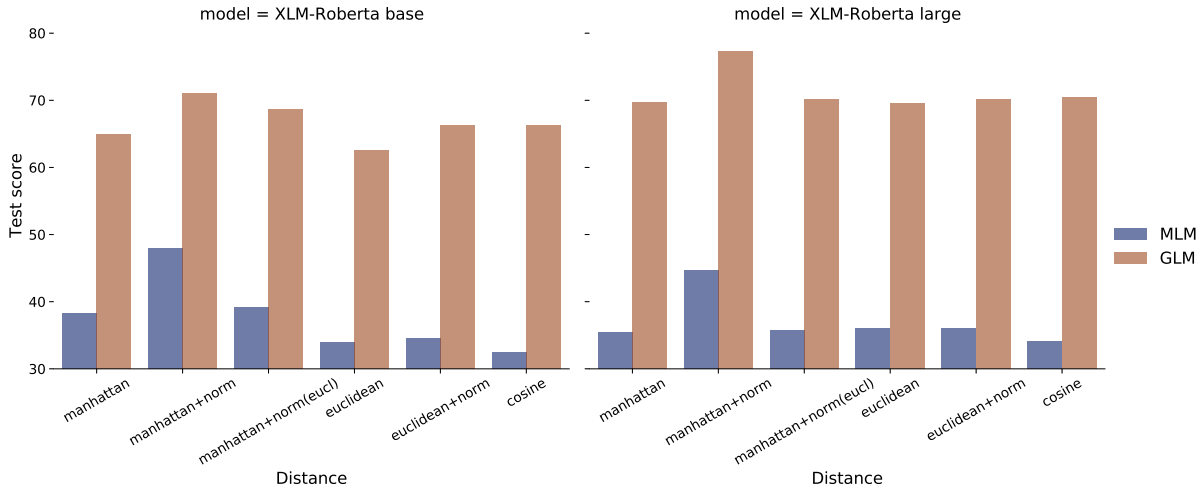

Figure 1: Mean test Spearman score for the GLM and MLM models for each distance measure.

representation:

$$r_{w_c} = T_c(c)[i]$$

For target words that are tokenized into multiple subword units, we average representations of these subwords.

Gloss encoder (T_g) takes as input a gloss $g_s = g_0, g_1, ..., g_m$ that defines a word sense s and encodes it as:

$$r_s = T_g(g_s)[0]$$

Taking the output from the first input token, which should be [CLS] or <s> token.

We can score each of the possible senses $s \in S_w$, for a target word w_c by taking the dot product of r_{w_c} against every r_s for $s \in S_w$:

$$\phi(w_c, s) = r_{w_c}^T r_s$$

As there is no such big WSD dataset as SemCor [14] for non-English languages, we extend BEM [2] system to the multilingual setting by replacing BERT with XLM-RoBERTa model [13] and exploiting its zero-shot cross-lingual transferability.

Both encoders were initialized with XLM-R base or large weights. Then the whole system was pre-trained on WSD data with cross-entropy loss. We denote this pre-training procedure as Gloss Language Modeling (GLM). In our experiments, these encoder models were not fine-tuned on any Russian data.

3.2 Sentence sampling

Following the competition's recommendations, we exploited Russian National Corpus (RNC) to sample sentences from the given epochs. We used rulemma[3] lemmatizer to find all occurrences of the target words in all forms in the corpus. For each target word, we sampled no more than 100 sentences per epoch.

3.3 Inter-epoch difference

In order to calculate desired $compare_{e_1,e_2}(w)$ value which denotes predicted compare metric for the word w in e_1, e_2 epochs pair, we build contextualized word representations from the sampled sentences and calculate an inter-corpus average distance between them. More precisely, we compute $d(w_{c_{e_1}}, w_{c_{e_2}})$ which is a distance between contextualized embeddings obtained from the context encoder of our WSD

[3] `https://github.com/Koziev/rulemma`

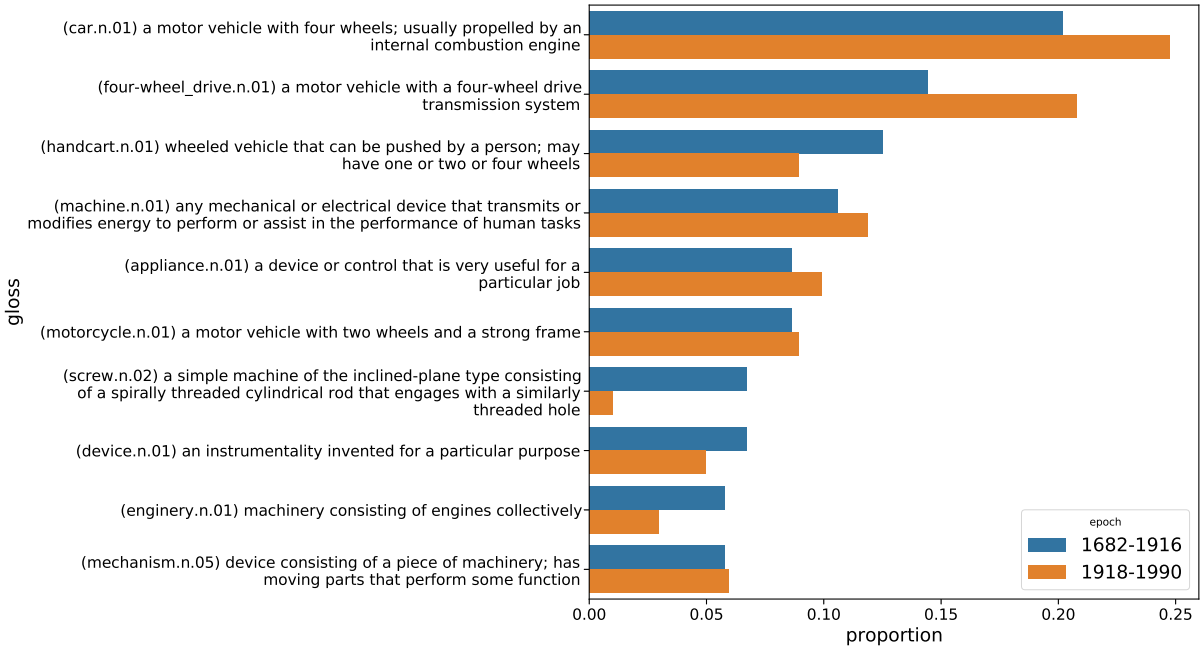

Figure 2: A proportion of examples with the word *машина* (car, vehicle, engine, computer) from the RuSemShift [11] samples where a particular gloss was selected (in top3) by GLM model. As this word did not occur in the post-Soviet part of RuSemShift [11], we show only the first two epochs.

pre-trained system for the word entries $w_{c_{e_1}}$ and $w_{c_{e_2}}$ of the word w from epochs e_1 and e_2 respectively. $compare_{e_1,e_2}(w)$ is calculated as average $d(\cdot)$ for all sampled sentence pairs $S_w(e_1, e_2)$ with the word w from the considered epochs pair e_1, e_2.

During the competition, we experimented on the $d(\cdot)$ definition based on our gloss-informed models. Here we propose methods that fully rely on the Context encoder and thus do not require any additional vocabulary or glosses. We achieve such generalization by using only outputs from the trained Context encoder.

1. **Euclidian (L2)**: Euclidian distance between outputs of the encoder.
2. **Euclidian+norm**: Euclidian distance between L2 normalized outputs of the encoder.
3. **Manhattan (L1)**: Manhattan distance between outputs of the encoder.
4. **Manhattan+norm**: Manhattan distance between L1 normalized outputs of the encoder.
5. **Manhattan+norm(Eucl)**: Manhattan distance between L2 normalized outputs of the encoder.
6. **Dot product**: Dot product similarity between outputs of the encoder.
7. **Dot product+norm(Manh)**: Dot product similarity between L1 normalized outputs of the encoder.
8. **Cosine**: Cosine similarity between outputs of the encoder.

As the bigger L1 or L2 distance means the bigger semantic change, to get the positive Spearman correlations with the gold scores we inverted our final average distances.

Instead of a single distance function, we can use a trainable combination of them. During the competition, we trained several regression models with features taken from 7 different distance measures: L1, L1+norm, L2, L2+norm, Dot product, Dot product+norm(Manh), Cosine. The models were trained to approximate human scores for the RuSemShift [11] sentence pairs.

Besides comparing distance measures, we also experimented on the encoders' initialization. In the result section, we compare the performance of the models, initialized with XLM-R base and XLM-R large [13].

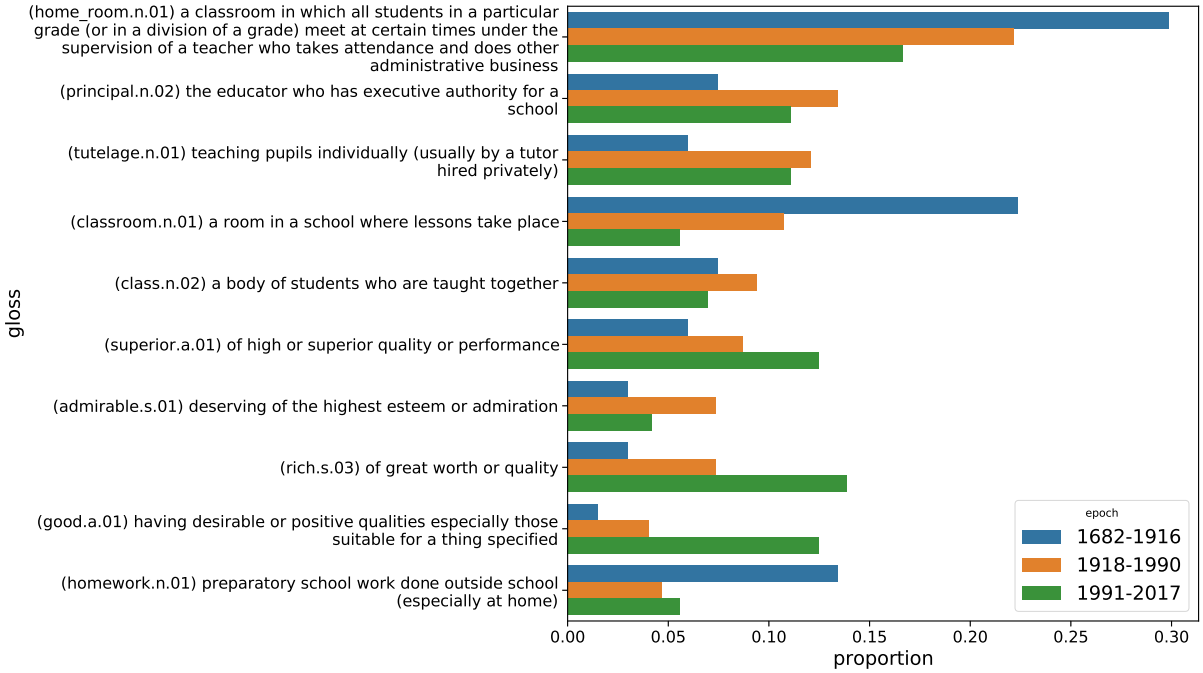

Figure 3: A proportion of examples with the word *классный* (classroom, cool, classy) from the RuSemShift [11] samples where a particular gloss was selected (in top3) by GLM model.

4 Experiments and Results

We trained our models on English SemCor [14] with glosses from WordNet 3.0 [8]. Systems based on XLM-RoBERTa base and XLM-RoBERTa large [13] were trained 20 and 10 epochs respectively. The Context and Gloss encoders were optimized on separate V100 GPUs for about 3 days for each backbone. Following standard practices, we used SemEval-2007 [12] as our development set to choose the final checkpoints. We evaluated our systems with the WSD framework proposed in [9].

4.1 GLM vs MLM

Figure 1 shows the gap between MLM and GLM pre-training, where we use distances between contextualized word embeddings obtained from the Context encoder pre-trained to solve the WSD task. Experiments show that models trained with GLM procedure strongly outperform their MLM counterparts regardless of distance measure or backbone. We also see that for both backbones manhattan+norm distance performs the best. In addition, the figure shows that GLM large model outperforms the base counter-part, but with MLM pre-training the base model performs slightly better.

4.2 Submissions

Table 1 shows overall results for the competition's test set for each of our submissions.

5 Qualitative analysis

As our GLM models were pre-trained to choose a proper definition for a target word, it is natural to try to interpret some of the models predictions. With the gloss knowledge from WordNet [8] we tried to find out how the meanings of the words were changing through time. We took the set of considered words from the book [3], where authors described the history of 20 Russian words. For each considered word w we run the following algorithm:

1. We sample all sentences with the target word w from RuSemShift [11].

Figure 4: Meanings' shift charts for the words *машина* (car, vehicle, engine, computer) and *классный* (classroom, cool, classy) from the book [3]. We also provide brief English translations of the columns and the words' meanings.

2. For each of these sentences we retrieve 3 glosses from WordNet having the maximum dot product with the contextualized embedding of the target word.
3. For each gloss we can calculate a proportion of examples with the target word w where this particular gloss was selected (in top3) by GLM model.

Figures 2, 3 show the dynamic of the meanings' changes for the words *машина* (car, vehicle, engine, computer) and *классный* (classroom, cool, classy) respectively. We took a pre-trained model with XLM-R large backbone for this purpose.

As we can see from these figures our model can choose sensible English glosses even from all WordNet [8] synsets, although the target words and their contexts are in Russian. Moreover, for these particular words, we can see consistency with the charts from the book [3] (Figure 4).

However, this approach with decoding Russian word senses with English WordNet [8] has several limitations. We have seen one of them during experiments with the word *пионер* (pioneer, scout), which of course drastically changed its meaning in the Soviet epoch. The Soviet meaning of this word is strongly connected to the Communist ideology, but in English, the nearest concept for this Soviet meaning is *scout*, which of course doesn't mean exactly the same, and consequently interpretation model can not find the proper English gloss and this leads to poor performance on such words.

6 Conclusion

In this paper, we proposed training a Gloss Language Model (GLM) to obtain better contextualized embeddings for the Russian Semantic Change Detection task. We have shown that this training procedure greatly boosts the performance compared to the traditional embeddings from a Masked Language Model (MLM) regardless of the distance measure employed.

Apart from that, we proposed a technique for the interpretation and visualization of the semantic

changes through time by linking Russian word occurrences to the reasonable definitions from the English WordNet and comparing distributions over those definitions for each epoch. Also, we discussed the limitations of this algorithm due to the difficult-to-translate concepts.

Acknowledgments

This research was supported in part through computational resources of HPC facilities at NRU HSE.

References

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — Minneapolis, Minnesota : Association for Computational Linguistics, 2019. — Jun. — P. 4171–4186. — Access mode: https://www.aclweb.org/anthology/N19-1423.

[2] Blevins Terra, Zettlemoyer Luke. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders // Proceedings of the 58th Association for Computational Linguistics. — 2020. — Access mode: https://blvns.github.io/papers/acl2020.pdf.

[3] Daniehl' M. A., Dobrushina N. R. Dva veka v dvadtsati slovakh. — NRU Higher School of Economics Publ. House, 2016. — ISBN: 9785759811480.

[4] Laicher Severin, Kurtyigit Sinan, Schlechtweg Dominik et al. Explaining and Improving BERT Performance on Lexical Semantic Change Detection. — 2021. — 2103.07259.

[5] GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge / Luyao Huang, Chi Sun, Xipeng Qiu, Xuanjing Huang // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — Hong Kong, China : Association for Computational Linguistics, 2019. — Nov. — P. 3507–3512. — Access mode: https://www.aclweb.org/anthology/D19-1355.

[6] Kutuzov Andrey, Pivovarova Lidia. RuShiftEval: a shared task on semantic shift detection for Russian // Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference. — 2021.

[7] Lesk Michael. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone // Proceedings of the 5th Annual International Conference on Systems Documentation. — SIGDOC '86. — New York, NY, USA : Association for Computing Machinery, 1986. — P. 24–26. — Access mode: https://doi.org/10.1145/318723.318728.

[8] Miller George A. WordNet: A Lexical Database for English // Commun. ACM. — 1995. — Nov. — Vol. 38, no. 11. — P. 39–41. — Access mode: https://doi.org/10.1145/219717.219748.

[9] Raganato Alessandro, Camacho-Collados Jose, Navigli Roberto. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. — Valencia, Spain : Association for Computational Linguistics, 2017. — Apr. — P. 99–110. — Access mode: https://www.aclweb.org/anthology/E17-1010.

[10] Liu Yinhan, Ott Myle, Goyal Naman et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. — 2019. — 1907.11692.

[11] Rodina Julia, Kutuzov Andrey. RuSemShift: a dataset of historical lexical semantic change in Russian // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 1037–1047. — Access mode: https://www.aclweb.org/anthology/2020.coling-main.90.

[12] SemEval-2007 Task-17: English Lexical Sample, SRL and All Words / Sameer Pradhan, Edward Loper, Dmitriy Dligach, Martha Palmer // Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). — Prague, Czech Republic : Association for Computational Linguistics, 2007. — Jun. — P. 87–92. — Access mode: https://www.aclweb.org/anthology/S07-1016.

[13] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // arXiv preprint arXiv:1911.02116. — 2019.

[14] Using a Semantic Concordance for Sense Identification / George A. Miller, Martin Chodorow, Shari Landes et al. // Proceedings of the Workshop on Human Language Technology. — HLT '94. — USA : Association for Computational Linguistics, 1994. — P. 240–243. — Access mode: https://doi.org/10.3115/1075812.1075866.

[15] Zero-shot Word Sense Disambiguation using Sense Definition Embeddings / Sawan Kumar, Sharmistha Jat, Karan Saxena, Partha Talukdar // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 5670–5681. — Access mode: https://www.aclweb.org/anthology/P19-1568.