DOI: 10.15514/ISPRAS-2021-33(6)-13

Weakly Supervised Word Sense Disambiguation Using Automatically Labelled Collections

[1] *A.S. Bolshina, ORCID: 0000-0002-9106-7192<angelina_ku@mail.ru>*
[2] *N.V. Loukachevitch, ORCID: 0000-0002-1883-4121<louk_nat@mail.ru>*
[1] *Lomonosov Moscow State University,*
GSP-1, Leninskie Gory, Moscow, 119991, Russia
[2] *Research Computing Center Lomonosov Moscow State University,*
GSP-1, Leninskie Gory, Moscow, 119991, Russia

Abstract. State-of-the-art supervised word sense disambiguation models require large sense-tagged training sets. However, many low-resource languages, including Russian, lack such a large amount of data. To cope with the knowledge acquisition bottleneck in Russian, we first utilized the method based on the concept of monosemous relatives to automatically generate a labelled training collection. We then introduce three weakly supervised models trained on this synthetic data. Our work builds upon the bootstrapping approach: relying on this seed of tagged instances, the ensemble of the classifiers is used to label samples from unannotated corpora. Along with this method, different techniques were exploited to augment the new training examples. We show the simple bootstrapping approach based on the ensemble of weakly supervised models can already produce an improvement over the initial word sense disambiguation models.

Keywords: Word sense disambiguation; Russian dataset; RuWordNet.

For citation: Bolshina A.S., Loukachevitch N.V. Weakly supervised word sense disambiguation using automatically labelled collections. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 6, 2021, pp. 193-204 (in Russian). DOI: 10.15514/ISPRAS–2021–33(6)–13

Acknowledgements. This research has been supported by the Interdisciplinary Scientific and Educational School of Lomonosov Moscow State University “Brain, Cognitive Systems, Artificial Intelligence”.

Разрешение неоднозначности на основе псевдоаннотированной коллекции

[1] *А.С. Большина*, *ORCID: 0000-0002-9106-7192 <angelina_ku@mail.ru>*
[2] *Н.В. Лукашевич, ORCID: 0000-0002-1883-4121 <louk_nat@mail.ru>*
[1] *Московский государственный университет имени М.В. Ломоносова,*
119991, Россия, Москва, Ленинские горы, д. 1
[2] *Научно-исследовательский вычислительный центр МГУ,*
119991, Россия, Москва, Ленинские горы, д. 1., стр. 4

Аннотация. Передовые системы разрешения неоднозначности основаны на обучении с учителем, однако для создания таких моделей требуются большие объемы размеченных данных, которые отсутствуют для большинства языков с ограниченными ресурсами. Для того, чтобы решить проблему недостатка аннотированных данных в русском языке, в данной статье предлагается подход для автоматической разметки значений многозначных слов с использованием ансамбля моделей, базирующихся на слабо контролируемом обучении. Для первичной разметки данных использовался автоматический метод, основанный на концепте однозначных родственных слов. С помощью этих синтетических данных были обучены три модели для разрешения неоднозначности, которые затем применялись в ансамбле для получения значений ключевых многозначных слов. Проведенные

эксперименты показали, что модели, обученные на данных, размеченных предобученными моделями, демонстрируют более высокое качество разрешения неоднозначности. Помимо этого, в статье изучается влияние различных подходов к аугментации текстовых данных на качество предсказаний.

Ключевые слова: автоматическое разрешение неоднозначности; датасеты на русском языке; RuWordNet

Для цитирования: Большина А.С., Лукашевич Н.В. Разрешение неоднозначности на основе псевдоаннотированной коллекции. Труды ИСП РАН, том 33, вып. 6, 2021 г., стр. 193-204 (на английском языке). DOI: 10.15514/ISPRAS–2021–33(6)–13

Благодарности: Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского государственного университета имени М.В. Ломоносова «Мозг, когнитивные системы, искусственный интеллект».

1. Introduction

The task of Word Sense Disambiguation (WSD) consists in identifying the correct sense of a polysemous word in the context. As with many other NLP tasks, WSD suffers from the problem that is called the knowledge acquisition bottleneck. The recent advances in the field of WSD can be applied only to some languages because obtaining hand-crafted sense-labelled training collections is very expensive in terms of time and extensive human efforts. The low-resource languages do not have access to the large labelled collections that are necessary for training current state-of-the-art supervised models. And that hinders the development of the different applications closely related to the WSD task, for example, semantic text analysis, knowledge graph construction, machine translation, question answering, etc.

In recent years to address these challenges, practitioners turn to weak supervision that implies training models using data with imperfect labels, that can be obtained with some user-defined heuristics, external knowledge bases, other classifiers etc. Various methods of automatic acquisition of training samples have been invented in the field of WSD. In our research we utilize the method to automatically generate and label training collections with the help of monosemous relatives, that is a set of unambiguous words (or phrases) related to particular senses of a polysemous word. The labels obtained with the help of this approach were used to train three different weakly supervised WSD models: logistic regression with the deep representations from ELMo [1] language model as features, fine-tuned BERT [2] model and BERT model trained on context-gloss pairs.

In this article, we propose an algorithm based on the ensemble of weakly supervised WSD models that can be used to label raw texts and, thus, reduce human efforts to annotation. The additional data provided by the algorithm was used to re-train original models, and the experiments showed that it enhanced the initial models’ performance. Moreover, leveraging different augmentation techniques we were also able to improve upon the classification results.

The paper is organized as follows. In section 2 we review the related work. Section 3 is devoted to the data description. The fourth section describes the method applied to automatically generate and annotate training collections. The models and augmentation techniques are presented in the fifth section. In the sixth section, we describe an algorithm based on the weighted probabilistic ensemble of the WSD models used to predict sense labels and in Section 7 we demonstrate the results obtained by three different models. Concluding remarks are provided in the eighth section.

2. Related work

To overcome the limitations, that are caused by the lack of annotated data, several methods of generating and harvesting large train sets have been developed. There exist many techniques based on different kinds of replacements, which do not require human resources for tagging. The most popular method is that of monosemous relatives [3]. Usually, WordNet [4] is used as a source for such relatives. WordNet is a lexical-semantic resource for the English language that contains a

description of nouns, verbs, adjectives, and adverbs in the form of semantic graphs. All words in those networks are grouped into sets of synonyms that are called synsets.

Monosemous relatives are those words or collocations that are related to the target ambiguous word through some connection in WordNet, but they have only one sense, i.e. belong only to one synset. Usually, synonyms are selected as relatives but in some works hypernyms and hyponyms are chosen [5]. Some researchers replace the target word with named entities [6], some researchers substitute it with meronyms and holonyms [7]. In the work [8] distant relatives (including distant hypernyms and hyponyms) were used; the procedure of training contexts selection was based on the distance to a target word and the type of the relation connecting the target sense and a monosemous relative.

Multilingual resources such as parallel corpora are also a valuable source of information that can be used to generate training collections for the WSD task [9]–[12]. Other methods of automatic annotation of training collections for WSD exploit knowledge bases like Wikipedia and Wiktionary [13]–[16]. In bootstrapping approach, the classifier relies on a small number of labelled seed instances, then a set of raw samples with the highest confidence is annotated using this model and utilized to retrain the model. The whole cycle of this procedure is repeated until the desired number of samples is labelled or some benchmark in performance is reached [17]–[19].

It is clear, that the above-mentioned methods cannot guarantee correct labelling of the samples, however, such imperfect data can still be used in weak supervision. This strategy is used extensively for named entity recognition [20], relation extraction [21], [22], entity linking [23] and text classification [24]. As weak supervision can introduce different types of noise into a model, in our research to infer the sense label of the unannotated sample, we combined the predicted class probabilities of the three weakly supervised models alongside uncertainty estimation.

Nowadays, the greater part of the WSD systems is based on neural networks. Recent studies have shown the effectiveness of contextualized word representations for the WSD task [25], [26]. The most widely used deep contextualized embeddings are ELMo [1] and BERT [2]. Let us briefly overview some of the state-of-the-art approaches in WSD. The system based on Transformer encoders, BERT contextualized word embeddings and sense vocabulary compression methods were introduced in [27]. The most significant feature of the algorithm EWISE is "predicting over a continuous sense embedding space as opposed to a discrete label space" [28]. The system EWISER [29] builds upon EWISE. But to better predict unseen words, the information about concepts and their relations from WordNet was added to this neural architecture. The work [30] focuses on exploring sparse contextualized word representations as a solution to the task of fine-grained WSD. The work [31] introduces the bi-encoder model for WSD. Context and gloss encoders are independent of each other and are initialized with BERT, but their output representations are combined to predict sense labels.

In the current research, we implemented a simple logistic regression model with ELMo representations as features, also we employed BERT for fine-tuning and sentence pair classification task between the sentence with a target polysemous word and the gloss related to one of its senses. These models are described in more detail in Section 5. We used automatically generated training collections to train these classifiers. Then with the help of these models, we annotated unlabeled data with noise labels and used it to train new WSD models.

3. Data

In our research as an underlying semantic network, we exploit Russian thesaurus RuWordNet [32]. It is a semantic network for Russian that has a WordNet-like structure. In total it contains 111.5 thousand words and word combinations for the Russian language. RuWordNet was used to extract semantic relations (e.g., synonymy, hyponymy etc.) between a target sense of a polysemous word and all the words (or phrases) connected to it, including those linked via distant paths. The sense inventory was also taken from this resource. RuWordNet contains 29297 synsets for nouns, 63014 monosemous and 5892 polysemous nouns. In this research, we consider only ambiguous nouns.

We utilized two types of corpora in the research. A news corpus consists of news articles harvested from various news sources. The texts have been cleaned from HTML elements or any markup. Another corpus consists of the several segments of Taiga corpus [33], which are compiled from news articles: Lenta.ru, Interfax, Komsomolskaya Pravda, Russian Magazines Hall, Fontanka.ru. We exploit these two corpora for extracting the sentences with target polysemous words and subsequent labelling by the ensemble of models. Moreover, the news corpus was exploited for the training word2vec model necessary for the algorithm of automatic generation of training collections.

Table 1. Quantitative characteristics of the target polysemous words and their senses

Target word and its sense	Number of validation samples	Number of glosses
аниматор$_0$ "a cartoonist"	29	11
аниматор$_1$ "an entertainer"	28	4
барометр$_0$ "a barometer"	24	17
барометр$_1$ "an indicator"	24	5
болячка$_0$ "an illness"	21	9
болячка$_1$ "a wound"	21	9
графит$_0$ "graphite"	24	5
графит$_1$ "a pencil"	17	6
дичь$_0$ "a fowl"	31	7
дичь$_1$ "nonsense"	17	9
зайчик$_0$ "sunbeam"	11	6
зайчик$_1$ "a bunny"	14	5
зародыш$_0$ "an embryo"	18	6
зародыш$_1$ "beginning"	10	7
калейдоскоп$_0$ "in the thick of it"	12	6
калейдоскоп$_1$ "kaleidoscope"	14	6
колыбель$_0$ "a crib"	16	5
колыбель$_1$ "the place of origin"	13	6
колокольчик$_0$ "a bluebell"	15	5
колокольчик$_1$ "a bell"	16	7

There are two variants of the WSD task: lexical sample and all-words. The former consists in disambiguating a small pre-selected set of polysemous words, the latter, on the contrary, implies predicting sense for each polysemous word in a text. In our experiments, we perform lexical sample sense disambiguation, which is why we chose several polysemous words, for which the WSD models of different types would be implemented, in advance. The selection criterion was as follows: the word should occur in the corpora at least 500 times and in no less than 200 documents. To evaluate the models, for each sense of the target polysemous words, we manually labelled small validation datasets compiled from the news articles from Wikinews.

Glosses are widely used in the field of WSD: for example, they can be utilized directly to disambiguate a sample [34] or can serve as a weakly supervised signal in models [31], [35]–[37]. The work [38] demonstrated that word definitions and examples of use can be used to augment training data and even boost the performance of a WSD system. For that reason, for each sense of a target polysemous word, we collected a small set of dictionary definitions and examples taken from dictionary entries. This data is intended to be added to a training collection along with the texts labelled by the ensemble of WSD models, and ultimately utilized in retraining.

The list of the selected polysemous words, the number of annotated samples and glosses for each sense is given in Table 1.

4. Method of automatic labelling of training collections

For the preliminary annotation of the training data, we employed the method described in [39], that is based on the concept of monosemous relatives. This approach for collecting a training corpus is based on the substitution: for every polysemous word we select appropriate monosemous relatives, then in a text, the occurrences of these relatives are substituted by the target polysemous word and these instances are labelled with a sense tag of a monosemous relative.

The findings from the research [39] showed, that the utilization of distant relatives (e.g., cohyponyms) along with synonyms, hyponyms and hypernyms enables a wider coverage of the target polysemous words in a training collection. In our research, the distance between the target sense of the polysemous word and its candidate monosemous relatives can reach up to four steps in the semantic graph.

Not all monosemous relatives are suitable as a representation of a target word polysemous word. To ensure, that the contexts with monosemous relatives extracted from a corpus will serve as good training samples for the target sense, we utilized a custom word2vec embedding model trained on the same corpus from which the contexts are retrieved. With the help of this model, we compute the similarity between the contexts, in which the candidate monosemous relative occur, and the words located close to the target polysemous word (within two steps from the target word).

For example, the selected monosemous relatives for the word болячка$_1$ “a wound” are "ссадина, волдырь, мозоль, оспина, прыщ, прыщик, струп" *(abrasion, blister, callus, pockmark, pimple, little pimple, scab*). For the word болячка$_0$ “an illness” the monosemous realtives are as follows: "болезнь, ангина, диабет, воспаление, бронхит, обморок, травматизм, астма, артроз" (*disease, sore throat, diabetes, inflammation, bronchitis, fainting, injury, asthma, arthrosis).*

This approach has already been applied to the subset of words from the RUSSE’2018 [40] evaluation dataset. The experiments showed that the models trained on the automatically generated collections can obtain the quality of disambiguation comparable to the models trained on the manually labelled data. In this research, we want to explore the application of the proposed method to other words with different level of ambiguity (see Section 3 for details).

5. Models

As it has already been said, in our research we employ three diverse supervised WSD models: two of them are based on ruBERT pretrained representations [41] released by DeepPavlov and the other employs RusVectōrēs [42] ELMo model trained on lemmatized Tayga corpus. The first model is a fine-tuned BERT with a sequence classification head: a linear layer on top of the concatenated target token representations from the last four hidden layers of the pre-trained transformer.

The second model is based upon the ideas from [36] and [43]: we utilized context-gloss pair with weak supervision for sentence pair classification task performed by the BERT model. The first element of a pair is a sentence with a target polysemous word; in each sentence, we put the target word in quotation marks as a weak supervised signal. The second element of a pair is a gloss definition of one of the senses of the ambiguous word. At the beginning of each gloss, we put the target word as a weak supervised signal. The two parts of training samples are concatenated with the special BERT symbol *[SEP]*, and are marked as positive ones only if the definition corresponds to the correct sense. Lemmatized context-gloss pairs from the automatically generated training collection are presented in Table 2. The experiments with Russian WSD models [25], [38] demonstrated that lemmatized training improves the performance of the models. For that reason, in all our models, we used lemmatized training samples, as can clearly be seen from Table 2.

Following [25], we also utilized in the experiments a simple logistic regression classifier, that uses ELMo representations as features. Additionally, [38] showed that the optimal way to use RusVectōrēs ELMo embeddings for the WSD task is to extract embedding solely for a target polysemous word, thus, for our experiments, we extracted the single vector of the target word from the ELMo top layer.

It should be noted, that in our research we investigate the performance of the monolingual WSD models trained on the automatically generated training collections and on the pseudo-labelled ones, that is why we do not explore multilingual WSD models. However, the results obtained in the described experiments can be used as a baseline of comparison for future work.

Table 2. Automatically created and labelled context-gloss pairs

Training sample	**Label**
кожа покраснение "болячка" припухлость круг глаз воспользоваться консилер [SEP] болячка : болезнь	0
кожа покраснение "болячка" припухлость круг глаз воспользоваться консилер [SEP] болячка : болезненный образование на тело	1
выключатель адреналин необходимый создание ингибитор "болячка" бета блокатор [SEP] болячка : болезненный образование на тело	0
выключатель адреналин необходимый создание ингибитор "болячка" бета блокатор [SEP] болячка : болезнь	1

6. Experimental design

Automatically generated data is noisy, and it is clear that models trained on such imperfect data may be prone to errors. In this article, we propose the solution to mitigate this problem common to weakly supervised systems. First, we train three models described in the previous section. Second, we utilize them to predict sense tag for each target sample in the unlabelled corpus. The core idea of our experiment is to use the ensemble of the three types of models to predict sense labels to the raw texts, taking into account the models’ uncertainty level. These texts would then constitute the new training dataset. Finally, all the above-mentioned classifiers would be retrained on this pseudo-annotated data.

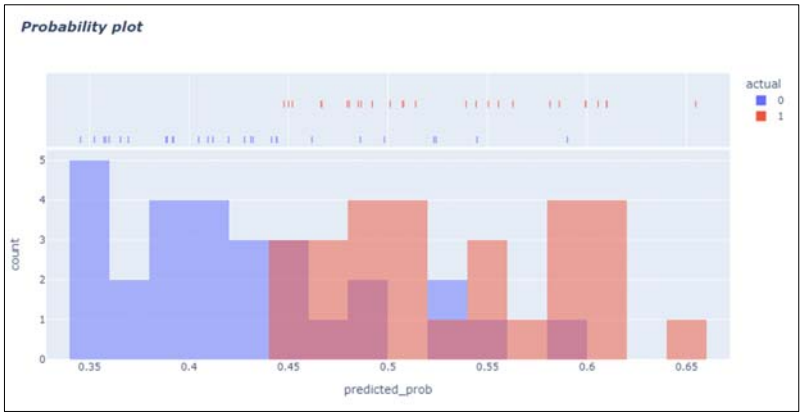

Fig. 1. The predicted probabilities for the validation samples with the word "аниматор": the logistic regression model

First of all, for each classifier, we defined the range of probabilities where it is uncertain. These thresholds help to filter poor classifiers’ predictions. A fine-tuned BERT model and a logistic regression classifier output a single probability of a class. We, therefore, employed the validation

dataset, and the models' probability estimations, in particular, to identify zones where the predictions get confused the most.

In Fig. 1, the area where the model makes most of the mistakes could be seen. Hence, in the case of this model, we would not trust the predictions that fall into the probability range from 0.45 to 0.6.

To predict the sense label with the help of the context-gloss pair model, one needs to compare the probabilities of all the context-gloss pairs available for this or that target word. To derive the criteria for this type of model, for each sense of the target word we analyzed the differences between the probabilities of the correct and incorrect class predicted for the validation samples. If this difference is more than 0 then the model predicted the right label. Thus, in our research, the 0.25-quantile of the positive values of difference is considered to be a threshold for the context-gloss pair model.

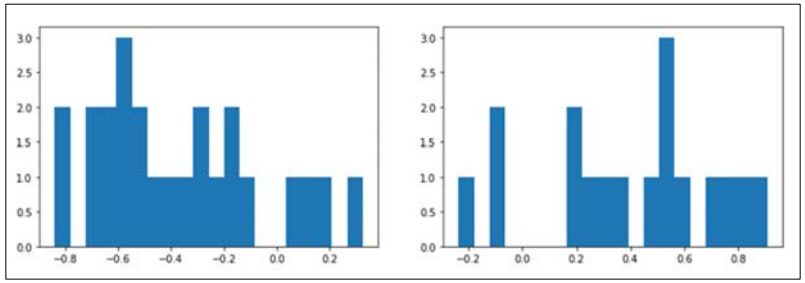

Fig. 2. The differences in the predicted probabilities for the validation samples with the word "графит": "graphite" and "a pencil" senses, respectively.

According to the data presented in Fig. 2, the threshold for the first class ("graphite") amounted to 0.1, whereas for the second class, this value is 0.3. So, if the difference between the classifier predictions is less than these values, we discard these probability values.

Thus, we have defined the threshold values for each of the WSD classifiers. The probability estimations, that does not meet the specified conditions, are discarded. To obtain the final class label from the probabilities, that comply with the requirements, we apply a weighting function to the models' output. There are various types of weighting schemes but we concentrate on the ones, where "rather than a single weight w_j, a separate weight is assigned to each class w_{ij}. This weight is set to be the proportion of cases correct for that class on the training data" [44]. Therefore, in our system, each base classifier prediction is multiplied by the precision value of this or that class obtained during the evaluation of the model on the validation set. Then all the weighted outcomes are summed, and the index of the maximum probability is returned as the final sense label for the sample with the target word. This weighting scheme allows us to rely on the predictions of other classifiers when some of them are not precise enough.

To boost the performance of the ensemble, we resorted to the principle "One sense per discourse" [45]: "if a polysemous word such as sentence appears two or more times in a well-written discourse, it is extremely likely that they will all share the same sense". From the news corpus and the segments of Tayga mentioned in Section 3, we extracted all the texts with the target polysemous words, in which they appeared more than twice. With the help of our scheme described above, for each occurrence of the target word, we predicted the class label. The final class label for all the samples, that constitute the text, is chosen according to the majority voting: we selected the class label that has more than half the votes. If this condition is not met, the label is determined by the maximum value of mean class probabilities.

Another constituent part of our experiment is augmentations. We have already described the augmentations with dictionary definitions and examples of use in Section 3. In addition, we applied easy data augmentation techniques from [46] to the samples labelled with the help of the ensembles. The maximum number of examples annotated by this strategy is 804 (total for the word "колокольчик"), which is not enough to train the models based on BERT. Consequently, we adapted

the original implementation of augmentation techniques for the WSD task in Russian: added the extraction of the synonyms from RuWordNet and imposed the limitation on the transformations of the original sentence (they should not involve the target polysemous word). The number of generated augmented sentences per original sentence was set to 6, this augmented data was used solely for retraining of the BERT-based models.

7. Results

In this section, we present the results of our experiments. It should be specially noted, that the evaluation on the validation dataset was performed with two different context windows: win=1 implies that the context includes one sentence before and after the sentence with the target word; when using win=0, we took only the sentence with the target word. In Table 3 we demonstrate the averaged f1-scores obtained with the models trained on the automatically generated training collections and the data pseudo-labelled by the ensemble of models. Table 4 contains the F1-scores for the models retrained on the new pseudo-annotated data with "One sense per discourse" assumption. In these tables, by (1) we denoted ELMo LogReg model, (2) is Fine-tuned BERT and (3) is Context-gloss pair BERT. We validated the models on the four variants of datasets: (a) is the dataset compiled without "One sense per discourse" principle and without dictionary definitions augmentation, (b) was composed without "One sense per discourse" principle but with dictionary definitions augmentation, (c) was created with "One sense per discourse" principle and without dictionary definitions augmentation, (d) was created with "One sense per discourse" principle and with dictionary definitions augmentation.

Table 3. Averaged classification results for the WSD models (F1-score)

Dataset	**ELMo LogReg**	**Fine-tuned BERT**	**Context-gloss pair BERT**
Dataset automatically labelled with the monosemous relatives approach	0.85	0.81	0.79
(a)	0.86	0.84	**0.87**
(b)	**0.86**	0.85	**0.86**
(c)	**0.87**	0.84	0.86
(d)	0.87	**0.88**	0.87

Table 4. Classification results for the WSD models trained on pseudo-labelled data with "One sense per discourse assumption" (F1-score).

Target word	**(c)**			**(d)**		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
аниматор win=1	0.74	0.7	0.78	0.79	**0.89**	0.85
аниматор win=0	0.8	0.7	0.77	0.79	0.8	0.77
барометр win=1	0.96	0.93	0.88	**0.98**	0.93	0.88
барометр win=0	0.94	0.91	0.87	0.92	0.94	0.9
болячка win=1	0.73	0.69	0.61	0.76	0.68	0.7
болячка win=0	0.76	0.69	0.74	**0.77**	0.73	0.75
графит win=1	0.6	0.59	0.78	0.65	**0.79**	0.74
графит win=0	0.65	0.59	0.72	0.63	0.74	0.74
дичь win=1	**0.97**	**0.97**	0.96	0.97	**0.97**	0.96
дичь win=0	0.9	0.9	0.95	0.87	**0.97**	0.95
зайчик win=1	**1**	**1**	0.96	**1**	**1**	0.96
зайчик win=0	**1**	**1**	0.98	**1**	**1**	0.98
зародыш win=1	0.95	**1**	0.96	0.95	**1**	0.96
зародыш win=0	0.86	0.95	0.88	0.86	0.95	0.92

калейдоскоп win=1	0.85	0.74	0.79	0.85	0.74	0.79
калейдоскоп win=0	**0.88**	0.74	0.84	**0.88**	0.77	0.87
колыбель win=1	**0.93**	0.87	0.86	**0.93**	0.87	0.89
колыбель win=0	0.9	0.9	**0.93**	0.9	0.9	**0.93**
колокольчик win=1	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**
колокольчик win=0	0.94	0.93	**0.97**	0.94	**0.97**	**0.97**
averaged f1 for win=1	0.87	0.85	0.86	**0.89**	**0.89**	0.87
averaged f1 for win=0	0.86	0.83	0.87	0.86	**0.88**	**0.88**

The results of the models retrained on the new texts labelled with the ensembles show that this procedure improves the overall performance of the models. In some cases, the retraining greatly increases the F1-score of a classifier, for example, the maximum f1-score 0.75 for the word "аниматор" (win=1) was achieved with the initial context-gloss pair BERT model, after the retraining on the pseudo-labelled data the score rose to 0.85. Sometimes the effect is less clearly defined, e.g. the results of the logistic regression model for the word "барометр" slightly change across different data modifications. However, sometimes we can see that the quality of the models trained on the pseudo-labelled data can be worse than the performance of the initial models: the LogReg classifiers trained for the word "болячка" on the data without using “One sense per discourse assumption” have lower f1-score than the initial classifiers.

In most cases, the models trained on the data that was obtained employing the “One sense per discourse” principle show higher results on the validation dataset. But there are several cases when this was not so: for example, the BERT models’ results for the word "калейдоскоп".

As for the dictionary definitions augmentations, the data shows that this additional data either has no effect on the performance score or, like in most of the cases, the f1-score improves. We also can see, that the window size has a varying effect on the performance of the WSD models, and its impact is yet to be investigated. Moreover, the data demonstrates that there is no clear trend in the type of the model that performs best: in some cases, ELMo has the highest f1-score ("барометр" (win=1), f1=0.98), sometimes it is fine-tuned BERT ("аниматор" (win=1), f1=0.89), context-gloss pair BERT outperformed all other models for the word "колыбель" (win=0), f1=0.93.

The experiments also proved that all the words are different in the degree of ambiguity. Some words have a very high f1-score, which means that they are easier to be disambiguated, for example, the word "зайчик", whose senses are well-differentiated. In contrast, the word "графит" has the metonymic type of the polysemy, its senses are connected as "the material-the product made of it". This word has a lower f1-score compared to the other words because its senses are hard to be identified.

The aim of our experiment is not to find the best WSD method in general. Rather, the goal is to find the method that improves the models trained on the data with weak labels. Our experiments proved that gradual retraining of the WSD models on the newly labelled data, i.e. bootstrapping, can enhance the overall performance of classification. Moreover, the proposed probabilistic ensemble weighting strategy can be utilized as an aid to manual sense annotation, for example, in the active learning environment.

8. Conclusion

In this work, we introduced the probabilistic ensemble weighting scheme, which is aimed at producing less noisy training data for the WSD classifiers. We proved that this strategy is robust in cases with automatically generated training collections, especially because we added an uncertainty estimation component. Moreover, additional data generated by the augmentation techniques have been shown to aid model performance.

The experiments demonstrated that retraining the models on the new data labelled utilizing the ensembles improved upon the initial results of the WSD models. The continuous retraining of the models on the new sets of samples can further boost the performance of lexical ambiguity resolution. Also, the probabilistic ensemble weighting scheme can be used to facilitate the efforts to manually label training data.

References

[1]. Peters M. E., Neumann M. et al. Deep contextualized word representations. In Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 2227–2237.

[2]. Devlin J., Chang M.-W. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2019, pp. 4171-4186.

[3]. Leacock C., Chodorow M., Miller G.A. Using corpus statistics and WordNet relations for sense identification. Computational Linguistics, vol. 24, no. 1, 1998, pp. 147–165.

[4]. Miller G. A. WordNet: a lexical database for English. Communications of the ACM, vol. 38, no. 11, 1995, pp. 39–41.

[5]. Przybyła P. How big is big enough? Unsupervised word sense disambiguation using a very large corpus. arXiv preprint arXiv:1710.07960, 2017.

[6]. Mihalcea R., Moldovan D.I. An Iterative Approach to Word Sense Disambiguation. In Proc. of the Thirteenth International Florida Artificial Intelligence Research Symposium Conference, 2000, pp. 219-223.

[7]. Yuret D. KU: Word sense disambiguation by substitution. In Proc. of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), 2007, pp. 207-214.

[8]. Martinez D., Agirre E., Wang X. Word relatives in context for word sense disambiguation. In Proc.of the Australasian Language Technology Workshop, 2006, pp. 42-50.

[9]. Taghipour K., Ng H.T. One million sense-tagged instances for word sense disambiguation and induction. In Proc. of the Nineteenth Conference on Computational Natural Language Learning, 2015, pp. 338–344.

[10]. Otegi A., Aranberri N. et al. QTLeap WSD/NED corpora: Semantic annotation of parallel corpora in six languages. In Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 2016, pp. 3023–3030.

[11]. Bovi C.D., Camacho-Collados J. et al. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (vol. 2: Short Papers), 2017, pp. 594–600.

[12]. Hauer B., Kondrak G. et al. Semi-Supervised and Unsupervised Sense Annotation via Translations. arXiv preprint arXiv:2106.06462, 2021.

[13]. Henrich V., Hinrichs E., Vodolazova T. WebCAGe–A Web-harvested corpus annotated with GermaNet senses. In Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 387-396.

[14]. Saif A., Omar N. et al. Building Sense Tagged Corpus Using Wikipedia for Supervised Word Sense Disambiguation. Procedia Computer Science, vol. 123, 2018, pp. 403-412.

[15]. Raganato A., Bovi C.D., Navigli R. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In Proc. of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 2894-2900.

[16]. Scarlini B., Pasini T., Navigli R. Just “OneSeC” for producing multilingual sense-annotated data. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 699-709.

[17]. Mihalcea R. Co-training and self-training for word sense disambiguation. In Proc. of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004), 2004, pp. 33-40.
[18]. Pham T.P., Ng H.T., Lee W.S. Word sense disambiguation with semi-supervised learning. Lecture Notes in Computer Science, vol. 3406, 2005, pp. 238-241.
[19]. Khapra M. M., Joshi S. et al. Together we can: Bilingual bootstrapping for WSD. In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 561-569.
[20]. Lison P., Barnes J., Hubin A. skweak: Weak supervision made easy for NLP. In Proc. of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, 2021, pp. 337-346.
[21]. Lin Y., Shen S. et al. Neural relation extraction with selective attention over instances. In Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), 2016, pp. 2124-2133.
[22]. Li Z., Hu F. et al. Selective kernel networks for weakly supervised relation extraction. CAAI Transactions on Intelligence Technology, vol. 6, no. 2, 2021, pp. 224-234.
[23]. Le P., Titov I. Boosting entity linking performance by leveraging unlabeled documents. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1935-1945.
[24]. Wang Y., Sohn S. et al. A clinical text classification paradigm using weak supervision and deep representation. BMC medical informatics and decision making, vol. 19, no. 1, 2019, pp. 1-13.
[25]. Kutuzov A., Kuzmenko E. To lemmatize or not to lemmatize: how word normalisation affects ELMo performance in word sense disambiguation. arXiv preprint arXiv:1909.03135, 2019.
- Wiedemann G., Remus S. et al. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In Proc. of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, 2019, pp. 161-170.

[26]. Vial L., Lecouteux B., Schwab D. Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. arXiv preprint arXiv:1905.05677, 2019.
[27]. Kumar S., Jat S. et al. Zero-shot word sense disambiguation using sense definition embeddings. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5670-5681.
[28]. Bevilacqua M., Navigli R. Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information. In Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2854-2864.
[29]. Berend G. Sparsity Makes Sense: Word Sense Disambiguation Using Sparse Contextualized Word Representations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8498-8508.
[30]. Blevins T., Zettlemoyer L. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. arXiv preprint arXiv:2005.02590, 2020.
[31]. Loukachevitch N. V., Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V. Creating Russian wordnet by conversion. In Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference Dialogue, 2016, pp. 405-415.
[32]. Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser. In Proc. of "CORPORA-2017" International Conference, 2017, pp. 78-84.
[33]. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proc. of the 5th Annual International Conference on Systems Documentation, 1986, pp. 24-26.
[34]. Luo F., Liu T. et al. Incorporating glosses into neural word sense disambiguation. arXiv preprint arXiv:1805.08028, 2018.
[35]. Huang L., Sun C. et al. GlossBERT: BERT for word sense disambiguation with gloss knowledge. arXiv preprint arXiv:1908.07245, 2019.
[36]. Loureiro D., Jorge A. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. arXiv preprint arXiv:1906.10007, 2019.
[37]. Bolshina A., Loukachevitch N. Exploring the limits of word sense disambiguation for Russian using automatically labelled collections. In Proc. of the Linguistic Forum 2020: Language and Artificial Intelligence (LFLAI), 2020, 14 p.
[38]. Bolshina A., Loukachevitch N. Generating training data for word sense disambiguation in Russian. In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", 2020, pp. 119-132.

[39]. Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Arefyev N., Leontyev A., Loukachevitch N. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", 2018, pp. 547–564.
[40]. Kuratov Y., Arkhipov M.Y. Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv preprint arXiv: 1905.07213, 2019.
[41]. Kutuzov A., Kuzmenko E. WebVectors: a toolkit for building web interfaces for vector semantic models. Communications in Computer and Information Science, vol. 661, 2016, pp. 155-161.
[42]. Kohli H. Transfer learning and augmentation for word sense disambiguation. In Advances in Information Retrieval, Springer, 2021, pp. 303-311.
[43]. Large J., Lines J., Bagnall A. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. Data Mining and Knowledge Discovery, vol. 33, no. 6, 2019, pp. 1674-1709.
[44]. Gale W. A., Church K. W., Yarowsky D. One sense per discourse. In Proc. of the Workshop on Speech and Natural Language, 1992, pp. 233-237.
[45]. Wei J., Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proc. of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6383-6389.

Информация об авторах / Information about authors

Ангелина Сергеевна БОЛЬШИНА – аспирант кафедры теоретической и прикладной лингвистики. Сфера научных интересов: автоматическая обработка текстов, автоматическое разрешение неоднозначности, глубокое обучение.

Angelina Sergeevna BOLSHINA – PhD student. Research interests: natural language processing, word sense disambiguation, deep learning.

Наталья Валентиновна ЛУКАШЕВИЧ – доктор технических наук, ведущий научный сотрудник. Сфера научных интересов: автоматическая обработка текстов, онтологии, анализ тональности.

Natalia Valentinovna LOUKACHEVITCH – Doctor of Technical Sciences, leading researcher. Research interests: natural language processing, ontologies, sentiment analysis.