Создание псевдоаннотированного обучающего корпуса для задачи разрешения лексической неоднозначности с помощью ансамбля моделей

А. С. Большина[1]

В настоящее время для задачи разрешения лексической неоднозначности наилучшие результаты на стандартных бенчмарках показывают алгоритмы, которые основаны на обучении с учителем. Однако, использование больших объемов размеченных данных для обучения таких моделей ограничивает их применение для языков с малым количеством ресурсов. Для русского языка также актуальна проблема нехватки аннотированных данных. В данной работе исследуется метод для автоматической разметки текстов, который основан на ансамбле моделей, предварительно обученных на синтетических данных. Результаты экспериментов демонстрируют, что модели, обученные на данных, размеченных предобученными моделями, показывают более высокое качество разрешения неоднозначности.

Ключевые слова: автоматическое разрешение неоднозначности; датасеты на русском языке; ELMo; BERT.

1. Введение

Задача автоматического разрешения неоднозначности состоит в определении корректного значения многозначного слова в том или ином контексте. Как и во многих других областях автоматической обработки текстов, в сфере разрешения неоднозначности остро стоит проблема недостатка размеченных данных. Для получения аннотированного корпуса необходимого для обучения размера требуется очень много времени и трудозатрат, поэтому последние достижения в области предсказания значений слов могут быть применены на практике лишь к языкам с достаточным количеством ресурсов. Русский язык относится к языкам с ограниченными ресурсами ввиду того, что для него нет больших обучающих коллекций с разметкой значений.

[1] *Большина Ангелина Сергеевна* — аспирант каф. теоретической и прикладной лингвистики филологического ф-та МГУ, e-mail: angelina_ku@mail.ru.

Bolshina Angelina Sergeevna — Lomonosov Moscow State University PhD student, Philological Faculty, Department of Theoretical and Computational Linguistics, e-mail: angelina_ku@mail.ru.

В последнее время для решения проблемы нехватки аннотированных данных специалисты обращаются к такой парадигме обучения как *weak supervision*, которая подразумевает обучение моделей на данных с неточными, зашумленными метками. Чаще всего такую разметку получают с помощью вручную заданных эвристик, внешних баз знаний, предсказаний предобученных классификаторов и т.п. Для автоматической аннотации значений слов в настоящей работе будет применяться ансамбль моделей, предварительно обученных на данных с метками низкого качества, которые также были получены с помощью автоматического метода.

В данном исследовании для создания синтетического датасета для первичного обучения моделей использовался метод, базирующийся на концепте однозначных родственных слов. Однозначные «родственники» – это слова или словосочетания, имеющие только одно значение и связанные каким-либо отношением с многозначным словом в семантической сети. Все методы, использующие данный концепт, основаны на заменах: для каждого значения многозначного слова выделяются однозначные родственные слова, затем в корпусе данные слова заменяются на ключевое неоднозначное слово, и все полученные предложения размечаются в соответствии со значением однозначного родственного слова. Существуют различные системы автоматической генерации обучающих данных, использующие концепт однозначных родственных слов ([1],[2],[3]), в текущей работе применялся алгоритм, описанный в [3]. Для того, чтобы оценить качество полученной с помощью ансамбля разметки, сравнивалось значение F1-меры базовых моделей, обученных на коллекции, собранной с помощью метода однозначных родственных слов, и F1-мера моделей, дообученных на новых данных. Пример однозначных родственных слов для многозначного слова *барометр* в значении «индикатор» приведены ниже:

1) *отправная точка* (когипоним), *движущая сила* (когипоним), *точка отсчета* (когипоним), *значимость* (когипоним), *мерило* (когипоним в 2-х шагах) ....

Коллекция, собранная с помощью метода однозначных «родственников», использовалась для обучения трех моделей: две из них используют предобученную модель ruBERT [4] от DeepPavlov, а другая базируется на языковой модели ELMo от RusVectōrēs [5], обученной на корпусе «Тайга»[1]. Первая модель – это тонко настроенный (*fine-tuned*) ruBERT с выходным слоем, предназначенным для классификации последовательностей: линейный слой, получающий на вход конкатенированные представления ключевого слова с четырех последних слоев предобученного трансформера. Вторая модель (*context-gloss pair BERT*) базируется на идеях, описанных в [6] и [7]: с помощью модели ruBERT решалась задача

[1] `https://tatianashavrina.github.io/taiga_site/`

классификации пар предложений, которые были представлены контекстом с ключевым многозначным словом и словарной дефиницией одного из его значений. Также, как и в работе [8], в настоящем исследовании для предсказания значения слова применялась логистическая регрессия, использующая эмбеддинги ELMo в качестве признаков.

2. Результаты

Для предсказания значений ключевых слов в корпусе предобученные модели использовались в ансамбле, так как синтетические данные, полученные с помощью метода однозначных родственных слов, могут содержать ошибки, и, соответственно, вносить шум в модель.

В данном исследовании также учитывалась степень «уверенности» каждой из моделей. Вероятностные предсказания логистической регрессии и модели fine-tuned ruBERT на валидационном датасете анализировались для выявления областей, где модели совершают наибольшее число ошибок. Для модели context-gloss pair BERT для каждого значения целевого слова рассматривалась разница между вероятностями правильного и неправильного класса, предсказанными для примеров из валидационной выборки. Если эта разница больше 0, значит модель предсказала правильную метку. Таким образом, в данном исследовании 0,25-квантиль положительных значений разницы считался порогом «уверенности» для модели context-gloss pair BERT. Чтобы получить окончательную метку класса из вероятностных оценок моделей, к ним применялась весовая функция. В предлагаемой системе каждое предсказание базового классификатора умножается на значение точности того или иного класса, полученное в ходе оценки модели на валидационном датасете. Затем все взвешенные результаты суммируются, и индекс максимальной вероятности возвращается в качестве конечной метки смысла для примера с целевым словом. Эта схема взвешивания позволяет учитывать только предсказания классификаторов с высокой степенью «уверенности».

Стоит также отметить, что в экспериментах применялись различные варианты разметки текстов: без использования принципа “One sense per discourse” [9] (вариант (а) – без аугментации (добавления) данных словарными дефинициями и примерами употребления слов, (б) – с аугментацией) и с применением данного принципа ((в) – без аугментации, (г) – с аугментацией). Результаты, полученные моделями, обученными на различных датасетах, приведены в таблице 1.

Результаты моделей, переобученных на новых текстах, которые были размечены ансамблями, показывают, что эта процедура улучшает качество моделей разрешения неоднозначности.

Таблица 1. Усредненные значения F1-меры для всех ключевых многозначных слов на валидационном датасете.

Датасет	**ELMo LogReg**	**Fine-tuned BERT**	**Context-gloss pair BERT**
Датасет, размеченный с помощью метода однозначных родственных слов	0.85	0.81	0.79
(а)	0.86	0.84	**0.87**
(б)	**0.86**	0.85	**0.86**
(в)	**0.87**	0.84	0.86
(г)	0.87	**0.88**	0.87

Список литературы

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

[2] Mihalcea R., Moldovan D. I., "An Iterative Approach to Word Sense Disambiguation", *In Proceedings of FLAIRS Conference*, 2000, 219–223.

[3] Bolshina A., Loukachevitch N., "Generating training data for word sense disambiguation in Russian", *In Proceedings of Conference on Computational Linguistics and Intellectual Technologies Dialog-2020*, 2020, 119–132.

[4] Kuratov Y., Arkhipov M. Y., "Adaptation of deep bidirectional multilingual transformers for Russian language", *In Proceedings of the International Conference "Dialogue 2019"*, 2019, 333–339.

[5] Kutuzov A., Kuzmenko E., "WebVectors: a toolkit for building web interfaces for vector semantic models", *In International Conference on Analysis of Images, Social Networks and Texts*, 2016, 155—161.

[6] Huang L., Sun C., Qiu X., Huang X., "GlossBERT: BERT for word sense disambiguation with gloss knowledge", *arXiv preprint arXiv:1908.07245*, 2019.

[7] Kohli H., "Transfer learning and augmentation for word sense disambiguation", *Advances in Information Retrieval*, 2021, 303—311.

[8] 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.

[9] Gale W. A., Church K. W., Yarowsky D., "One sense per discourse", *USA: Association for Computational Linguistics*, 1992, 233—237.

The Creation of Pseudo-Annotated Data for Word Sense Disambiguation using Ensembles of Models

Angelina Bolshina

Nowadays, supervised word sense disambiguation (WSD) algorithms attain the best results on the main benchmarks. However, large sense-tagged training sets are required for their training. This requirement

hinders the development of the word sense disambiguation systems for many low-resource languages, including Russian. To address the issue of the knowledge acquisition bottleneck in Russian, in this work we investigate the method for automatic text labelling that is based on the ensemble of weakly supervised WSD models. Our experiments demonstrated that the models retrained on the new pseudo-annotated data outperform the initial models.

Keywords: Word sense disambiguation; Russian dataset; ELMo; BERT.

References

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

[2] Mihalcea R., Moldovan D. I., "An Iterative Approach to Word Sense Disambiguation", *In Proceedings of FLAIRS Conference*, 2000, 219–223.

[3] Bolshina A., Loukachevitch N., "Generating training data for word sense disambiguation in Russian", *In Proceedings of Conference on Computational Linguistics and Intellectual Technologies Dialog-2020*, 2020, 119–132.

[4] Kuratov Y., Arkhipov M. Y., "Adaptation of deep bidirectional multilingual transformers for Russian language", *In Proceedings of the International Conference "Dialogue 2019"*, 2019, 333–339.

[5] Kutuzov A., Kuzmenko E., "WebVectors: a toolkit for building web interfaces for vector semantic models", *In International Conference on Analysis of Images, Social Networks and Texts*, 2016, 155—161.

[6] Huang L., Sun C., Qiu X., Huang X., "GlossBERT: BERT for word sense disambiguation with gloss knowledge", *arXiv preprint arXiv:1908.07245*, 2019.

[7] Kohli H., "Transfer learning and augmentation for word sense disambiguation", *Advances in Information Retrieval*, 2021, 303—311.

[8] 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.

[9] Gale W. A., Church K. W., Yarowsky D., "One sense per discourse", *USA: Association for Computational Linguistics*, 1992, 233—237.