DOI: 10.15514/ISPRAS-2022-34(1)-10

Алгоритмы обработки естественного языка для понимания семантики текста

Д.О. Жаксыбаев, ORCID: 0000-0001-6355-5431 <darhan.03.92@mail.ru>
Г.Н. Мизамова, ORCID: 0000-0002-1012-9700 <mizamgul@mail.ru>
Западно-Казахстанский аграрно-технический университет имени Жангир хана
090009, Республика Казахстан, город Уральск, улица Жангир-хана, 51

Аннотация. Векторное представление слов используется для различных задач автоматической обработки естественного языка. Множество методов существует для представления векторов слов, включая методы нейронных сетей Word2Vec и GloVe, а также классический метод латентно-семантического анализа LSA. Цель данной работа посвящена исследованию эффективности использования сетевых векторных методов LSTM для неклассической классификации высоты тона в текстах на русском и английском языках. Описаны характеристики векторных методов классификации слов (LSA, Word2Vec, GloVe), описана архитектура нейросетевого классификатора слов на основе LSTM и взвешены методы векторной классификации слов, представлены результаты экспериментов, вычислительных средств и их обсуждение. Лучшей моделью векторного представления слов является модель Word2Vec, учитывая скорость обучения, меньший размер корпуса слов для обучения, большую точность и скорость обучения нейросетевого классификатора.

Ключевые слова: обработка теста; ключевые слова; процедура отбора; вектор слов

Для цитирования: Жаксыбаев Д.О., Мизамова Г.Н. Алгоритмы обработки естественного языка для понимания семантики текста. Труды ИСП РАН, том 34, вып. 1, 2022 г., стр. 141-150. DOI: 10.15514/ISPRAS–2022–34(1)–10

Natural Language Processing Algorithms for Understanding the Semantics of Text

D.O. Zhaxybayev, ORCID: 0000-0001-6355-5431 <darhan.03.92@mail.ru>
G.N. Mizamova, ORCID: 0000-0002-1012-9700 <mizamgul@mail.ru>
West Kazakhstan Agrarian and Technical University named after Zhangir Khan
51 Zhangir Khan Street, Uralsk, Republic of Kazakhstan, 090009

Abstract. Representation of vector words is used for various tasks of automatic processing of natural language. Many methods exist for the representation of vector words, including methods of neural networks Word2Vec and GloVe, as well as the classical method of latent semantic analysis LSA. The purpose of this paper is to investigate the effectiveness of using network vector methods LSTM for non-classical pitch classification in Russian and English texts. The characteristics of vector methods of word classification (LSA, Word2Vec, GloVe) are described, the architecture of neural network classifier based on LSTM is described and vector methods of word classification are weighted, the results of experiments, computational tools and their discussion are presented. The best model for vector word representation is Word2Vec model given the training speed, smaller word corpus size for training, greater accuracy and training speed of neural network classifier.

Keywords: test processing; keywords; selection procedure; word vector

For citation: Zhaxybayev D.O., Mizamova G.N. Natural Language Processing Algorithms for Understanding the Semantics of Text. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 1, 2022, pp. 141-150 (in Russian). DOI: 10.15514/ISPRAS–2022–34(1)–10

1. Введение

Обработка текста на естественном языке приобретает все большее значение и вызывает большой интерес в социологии, маркетинге, лингвистике, психологии и других областях человеческой деятельности. Быстрое распространение Интернета является одной из основных движущих сил этой тенденции.

Автоматизация обработки текстовой информации требует представления текста в виде числовой модели, которая представляет текст как вектор его характеристик – описание его атрибутов. чтобы получить

Метод представления символов (называемый текстовым индексированием) основан на двух подходах: представлении текста в виде «мешка слов» и представлении текста в виде последовательности слов, представленных в виде векторов [1].

Представляя текст в виде «мешка слов», текст можно представить в виде вектора словарного запаса обучающей выборки. Каждый элемент вектора представляет вес соответствующего слова в словаре. Весом может быть частота слова в тексте или какой-то другой более сложный показатель. Это представление не учитывает порядок слов в тексте, что является одним из основных недостатков этого метода.

Многообещающей тенденцией в обработке естественного языка является векторное представление слов, которое позволяет представить текст в виде последовательности векторов, соответствующих словам в тексте.

Простейшим способом преобразования слова в вектор является одиночное кодирование [2], то есть кодирование каждого слова вектором длины, равной размеру словаря обучающей выборки. Каждый из этих векторов состоит из нуля и единицы, соответствующей позиции слова в словаре. Это представление малоэффективно с точки зрения запоминания и, прежде всего, не дает никакого объяснения смыслового значения слов и не позволяет сравнивать слова по смысловой близости, что очень затрудняет классификацию.

Существует несколько методов получения векторных представлений слов, позволяющих создать низкоразмерное векторное представление для каждого слова в корпусе (наборе текстов) и сохранить контекстуальное сходство слов. Во-первых, известны два метода нейронных сетей: Word2Vec [3, 4], разработанный Google, и Global Vectors (GloVe) [5], разработанный в Стэнфордском университете.

Было показано, что эти два метода превосходно справляются с задачами обработки естественного языка, такими как распознавание похожих слов и т. д. Однако также было показано [5], что классические методы, в частности метод латентного семантического анализа LSA [6], могут быть более полезными, чем, например, Word2Vec. Это связано с тем, что Word2Vec с самого начала учится на низкоразмерных векторах и не использует всю информацию из учебного корпуса слов. Было показано [7], что метод LSA более надежен и существенно не зависит от размера корпуса.

Цель данной работы является изучение эффективности использования этих методов при автоматической обработке текстов на русском и английском языках. Обнаружение тона – это специфическая классификационная задача, которая включает в себя количественную оценку и категоризацию мнений, выраженных в тексте, с целью определения того, является ли отношение автора к той или иной теме, продукту положительным, отрицательным или нейтральным [5]. Искусственные нейронные сети (ИНС) оказались хорошим решением таких проблем. Архитектура используемых ИНС может различаться. ИНС прямого распространения [6] являются контекстно-зависимыми, то есть они могут учитывать только определенное количество слов вокруг данного слова, чтобы определить его значение повторно.

2. Методология исследования

В данной статье сравниваются три метода: LSA – классический метод, Word2Vec как наиболее популярный метод в русскоязычном сообществе среди программистов и GloVe как популярная альтернатива, мало описанная в русскоязычных источниках. Далее «слово» заменяется словом «терм», подчеркивая, что это текстовый элемент с определенной смысловой нагрузкой.

2.1 Семантический анализ

Метод латентного семантического анализа (LSA) основан на принципах факторного анализа [6], который позволяет, в том числе, выявить латентные (скрытые) отношения между терминами и текстами (документами), определяющие присущую документам тематику и термины. Каждая тема характеризуется своим значением в смысловом отношении документов и терминов. Размерность векторного представления уменьшена за счет исключения из лингвистической модели тем с наименьшей смысловой нагрузкой.

Первым шагом является преобразование основного текста (документов) в массив терминов. Элементами этой матрицы обычно являются веса TF-IDF [6].

TF (частотный термин) – это отношение nt (количества вхождений слова t) к общему количеству слов в документе:

$$\boldsymbol{TF(t,d)} = \frac{\boldsymbol{n_t}}{\sum_{\boldsymbol{k}} \boldsymbol{n_k}}.$$

Кроме того, согласно теореме о методах вычитания сингулярных разложений [7], полученную вещественную прямоугольную матрицу можно вычесть как произведение трех матриц:

$$A = USV^t,$$

где матрицы U и V ортогональны, а S – диагональная матрица, диагональ которой содержит сингулярные значения матрицы A.

Если в матрице S оставляем только наибольшие сингулярные значения k и матрицы U и V – только столбцы, соответствующие этим значениям, то произведение $\hat{S}$ и $\hat{V}$ матриц T есть наилучшее приближение исходной матрицы A к матрице $\hat{A}$ с рангом k.

Эта матрица представляет собой структуру ассоциативной зависимости, которая лежит в основе исходной матрицы, и каждый терм (строка в матрице $\hat{U}$) и каждый документ (строка в матрице $\hat{V}$) представлены в виде векторов в общем пространстве с размерностью k (пространство гипотез) (см. рис. 1). k подбирается опытным путем и зависит от количества исходных документов. Векторные термины можно использовать как векторные слова.

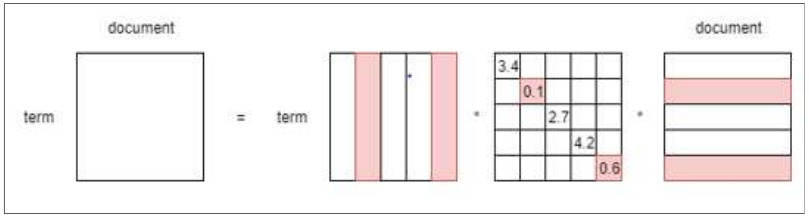

Рис. 1. Векторная модель LSA
Fig. 1. LSA vector model

2.2 Word2Vec

Word2Vec [3] –метод, позволяющий предсказывать контекст слова по заданному слову (метод Skip-Gram) или наоборот предсказывать слово по заданному контексту (метод CBOW). В этом случае скрытый слой нейронной сети проецирует слова на небольшой вектор.

Общая структура нейронной сети (см. рис. 2):

- входной слой, который получает векторы контекстных слов в случае CBOW или векторы одномерных кодирующих слов v в случае Skip-Gram;
- скрытый слой с несколькими нейронами k;
- выходной слой с иерархической функцией активации Softmax [4] или отрицательной выборкой [4], выход которой сходится во время обучения к векторам слов в случае CBOW или к векторам контекстных слов в случае Skip-Gram, в одноразмерное кодирование v.

После обучения выходные данные скрытого слоя используются для получения вектора слов размера k.

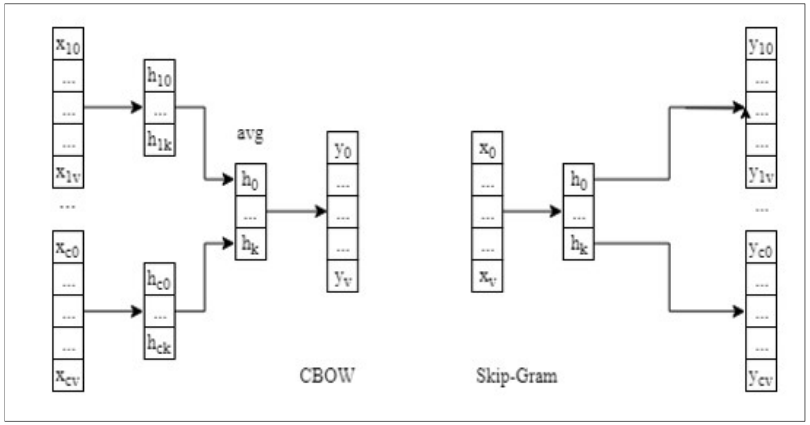

Рис. 2. Метод Word2Vec
Fig. 2. Word2Vec Method

Хотя этот метод охватывает только статистические свойства текстов, кажется, что обученная модель Word2Vec охватывает некоторые семантические свойства слов.

2.3 GloVe

GloVe [5] представляет собой модель, сочетающую в себе свойства метода уменьшения сингулярности и метода Word2Vec.

Первым шагом является создание матрицы состава X для тренировочного датасета. Значение элемента X_{ij} показывает, как часто слово j появляется в контексте слова i. Семантическая близость слов i и j оценивается с помощью отношения вероятности их совместного появления в контексте k:

$$F(w_i, w_j, \hat{w}_k) = \frac{P_{ik}}{P_{jk}} = \frac{X_{ik}/\sum_m X_{im}}{X_{jk}/\sum_n X_{jn}},$$

где w_i, w_j – векторы слов, $\hat{w}_k$ – вектор контекста.

Семантическая близость этих векторов определяется их скалярным произведением. С помощью преобразований и допущений в [5] показано, что целью формирования GloVe является изучение векторов таким образом, чтобы их скалярное произведение было близко к логарифму вероятности совпадения слов в выборке формирования. Чтобы уменьшить важность совпадений редких (менее информативных, шумных и ненадежных) или ненадежных слов, а также уменьшить важность очень распространенных (например, «есть») совпадений, Pennington и др.

В [5] в качестве целевой функции обучения (функции потерь) используется модель взвешенной регрессии по методу наименьших квадратов (1), где w_i – вектор

существительного, $\widehat{w}_j$ – вектор контекста, b_i и b_j – масштабирование значений отклонения слов существительного и контекста соответственно, V - размер словаря:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \hat{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

2.4 Классификация текстов

Основными этапами разработки и формирования текстового классификатора являются:

– предварительная обработка текста,
– индексация текста,
– разработка и формирование классификатора,
– оценка качества классификации.

Предварительная обработка текста обязательно предполагает токенизацию, т. е. выделение в тексте языковых единиц, так называемых лексем, слов или терминов.

Обработка текста может также включать

– сопоставление слов в верхнем и нижнем регистре, чтобы избежать семантических различий между похожими словами в разных регистрах;
– исключение таких слов, как союзы, предлоги, артикли, т.е. слова не похожие на слово, имеющие одинаковое значение, например, к ним относятся устранение семантически нейтральных слов (стоп-слов);
– удаление или замена цифр их текстовыми эквивалентами;
– устранение знаков препинания и лишних пробелов;
– нормализация слов: стемминг или лемматизация, замена всех слов стандартизированной формой.

Возможны и другие методы предварительной обработки слов. Выбор сценария предварительной обработки зависит от типа решаемой задачи и характеристик выборки. В результате выделяются все значимые слова, которые собственно и определяют смысловое значение текста.

Классификация текста требует индексации текста, т.е. описания каждого свойства текста, подлежащего классификации.

Всеобъемлющие и точные измерения, а также конкретные тестовые образцы могут использоваться для оценки качества классификации.

2.5 Определение тональности

Ставится задача сравнить эффективность методов представления векторов слов, рассмотренных в предыдущем разделе, при определении тональности текста.

Структура сети показана на рис. 3 и содержит следующие уровни.

1) Входной слой, который получает документ в виде последовательности лексемно-термальных индексов (метки).
2) Слой векторного словаря с фиксированными весами, где веса слоя последовательности определяются матрицей, строка которой i является векторным представлением термина i; используется для выбора соответствующего вектора термов; матрица строится из векторов, сгенерированных моделью векторного словаря.
3) Корректирующий слой, который изменяет определенный процент выходных значений предыдущего слоя, чтобы избежать чрезмерного обучения.
4) Сверточная сеть – это последовательность сверточных слоев и подвыборок с функцией максимума; он используется для уменьшения количества входных параметров

следующего слоя и для повышения скорости обучения.
5) Слой LSTM как основной классификатор.
6) Выходной слой с сигмоидальной функцией активации.

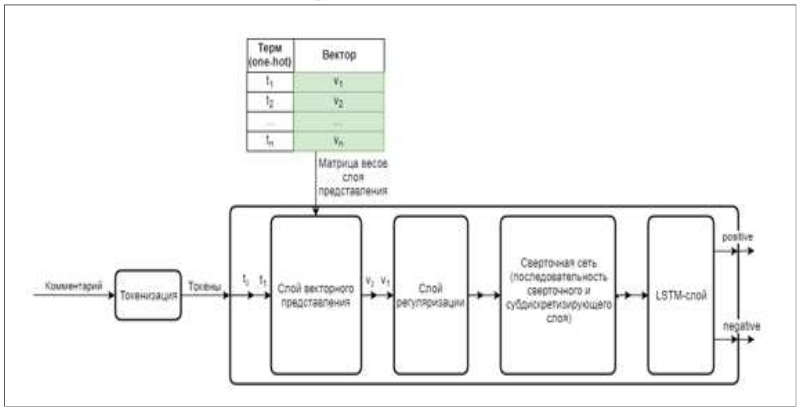

Рис. 3. Тональность текста
Fig. 3. Sentiment of the text

3. Результаты и обсуждения

В данной работе мы используем корпус обзоров социальных сетей на русском языке [4] и корпус обзоров фильмов на английском языке на IMDB [5].

Для каждого корпуса были проведены эксперименты по обучению разного количества текстов и с разными сочетаниями параметров вектора и шаблона слова:

– количество обучающих текстов – 1000, 10000, 25000;
– размерность векторного представления – 50, 150, 300;
– количество периодов или итераций обучения LSA – 2, 8, 15.

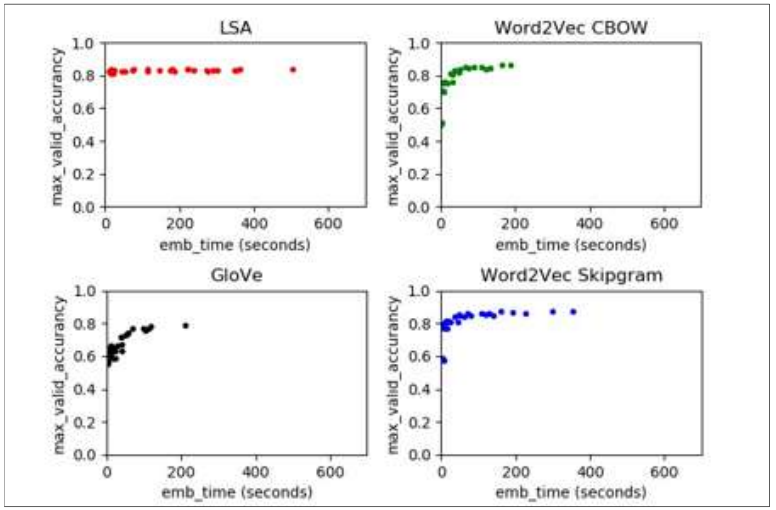

Рис. 4. Диаграммы рассеяния для англоязычного текста
Fig. 4. Scattering charts for English text

Для Word2Vec и GloVe использовался размер окна контекста: 2 слоя левее и 2 слова правее основного слова. Для Word2Vec использовался отрицательный размер выборки. Для каждого теста фиксировалось время обучения векторной модели слов (на рис. 4, 5 показан график по оси абсцисс). Чтобы проверить эффективность сети LSTM в определении тона текста, она была обучена 13 000 аннотаций. Проверки правдоподобия были выполнены для 3250 аннотаций. Поэтому максимальная точность классификации, полученная для тестовой выборки за 10 эпох обучения, была определена как максимально допустимая точность классификации (ось ординат на рис. 4, 5).

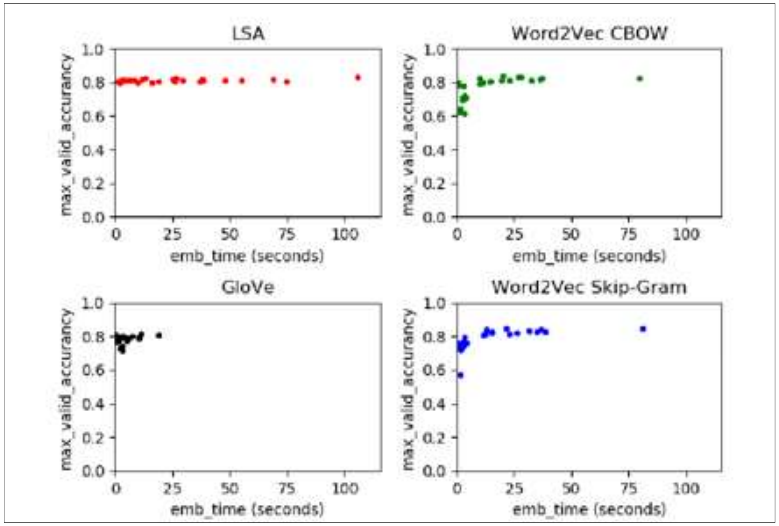

Рис. 4. Диаграммы рассеяния для русскоязычного текста
Fig. 4. Scattering charts for Russian text

У английских текстовых классификаторов были самые высокие средние значения точности (за исключением тестов, в которых использовался GloVe), но это может быть связано и с тем, что они были длиннее – обучающих данных было больше.

LSA является наиболее надежной моделью с точки зрения переменных параметров точности классификатора, так как все эксперименты дали примерно одинаковую точность классификации. LSA дал лучшие результаты по точности классификации (в среднем 81,2 % для русских текстов и 82,9 % для английских текстов), но в то же время самые плохие результаты по времени классификации как для русского, так и для английского языков. Широко используемые методы Word2Vec Skip-Gram и Word2Vec CBOW вели себя примерно одинаково для обоих корпусов: точность классификации снижалась при небольшом количестве учебных текстов. Скорость их обучения ниже, чем у моделей LSA, но значительно ниже, чем у моделей GloVe.

4. Заключение

Метод Word2Vec показал более высокую среднюю точность классификации и меньшую чувствительность к размеру обучающей выборки, несмотря на меньшую длину текста в русском корпусе. Эксперимент, в котором использовался этот метод, показал наивысшую точность. В случае с английским корпусом этот метод показал лучшие результаты по точности классификации и наименьшую чувствительность к размеру обучающей выборки.

Согласно [5], эксперты обычно сходятся во мнении о точности того или иного текста в 79% случаев. Эффективным считается классификатор, определяющий тональность текста с

точностью более 70%. В случае русского корпуса все классификаторы на основе LSA и GloVe могут быть использованы на практике по этому критерию (см. рис. 5). Как показывают вычислительные эксперименты, Word2Vec сложно применять к небольшим корпусам, но в других случаях он работает хорошо. В случае английского корпуса на практике можно использовать все классификаторы на основе LSA (см. рис. 4). Word2Vec также подходит для небольших корпусов, но в большинстве случаев работает хорошо. Большинство экспериментов с использованием модели GloVe показали неприемлемые результаты.

Обратите внимание, что для каждого набора параметров шаблонов представления вектора слов был выполнен только один тест, поэтому результаты велики.

Точность классификатора на основе сети LSTM также зависит от метода токенизации и обработки токенизации текста (обрезка текста или лемматизация, удаление стоп-слов, сокращение одиночного слова и т. д.), параметров векторного представления модели, а также структуру и параметры самой нейронной сети. Исследования, описанные в этой статье, были выполнены с ограниченным числом значений параметров для моделей представления векторов слов. Поэтому стоит рассмотреть различные способы обработки текстов и проверить эффективность моделей векторного представления с другими параметрами и другими сетевыми архитектурами LSTM, чтобы улучшить результаты. Стоит обратить внимание и исследовать эффективность GloVe.

Заслуживает внимания эффективность модели Word2Vec в других российских корпусах, так как она показывает очень хорошие результаты при анализе тонов. Однако примеров применения этой системы в русскоязычной литературе не обнаружено.

Выбор лучшей текстовой векторной модели также основывается на критериях классификации. Их критерии и приоритеты неодинаковы для разных задач, поэтому модель необходимо выбирать в соответствии с ограничениями, которым должна соответствовать выбранная модель.

Список литературы / References

[1] Chilakapati A. Word Bags vs Word Sequences for Text Classification. URL: https://towardsdatascience.com/word-bags-vs-word-sequences-for-text- classification-e0222c21d2ec, accessed 01.02.2022.
[2] Brownlee J. How to One Hot Encode Sequence Data in Python. URL: https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python accessed 05.02.2022.
[3] Le Q., Mikolov T. Distributed Representations of Sentences and Documents. In Proc. of the 31st International Conference on Machine Learning, 2014, pp. 1188-1196.
[4] Mikolov T., Chen K. et al. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781, 2013, 12p.
[5] Pennington J., Socher R., Manning C. GloVe: Global Vectors for Word Representation. In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532-1543.
[6] Landauer T.K., Foltz P.W., Laham D. Introduction to Latent Semantic Analysis. Discourse Processes, vol. 25, issue 2-3, 1998, pp. 259-284.
[7] Altszyler E., Sigman M., Slezak D.F. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. arXiv preprint arXiv:1610.01520, 14 p.

Информация об авторах/ Information about authors

Дархан Оракбаевич ЖАКСЫБАЕВ – магистр педагогических наук, окончил аспирантуру по специальности "Информационные системы" в Евразийском национальном университете имени Л. Гумилева, преподаватель кафедры информационных систем.

Darkhan Orakbayevich ZHAXYBAYEV – Master of Pedagogical Sciences, graduated from the PhD studies by the specialty "Information Systems" at the Eurasian National University named after L. Gumilyov, Lecturer of the Department of Information Systems.

Гулбаршын Нурлановна МИЗАМОВА – магистр технических наук, окончила аспирантуру по специальности "Информационные технологии и телекоммуникации", преподаватель кафедры информационных систем.

Gulbarshyn Nurlanovna MIZAMOVA – Master of Technical Sciences, graduated from the post-graduated studies in the specialty "Information technology and telecommunications" (PhD), Lecturer of the Department of Information.