

Средства автоматической классификации научных текстов в технологии ВИНТИ РАН

- ❑ Федорец Олег Владимирович, к.т.н., зав. лабораторией,
- ❑ Филимонов Алексей Викторович, главный специалист,
- ❑ Шапкин Александр Владимирович, к.т.н., главный специалист.

Классификационные схемы

- Входной поток более 1 млн документов в год.
 - Три стадии классификации потока:
 1. Разметка по 16 областям науки и техники:
Автоматика и радиоэлектроника, Астрономия, Биология, География, Геология, Информатика, Математика, Машиностроение, Metallургия, Механика, Охрана окружающей среды, Транспорт, Физика, Химия, Экономика промышленности, Электротехника.
 2. Распределение по выпускам Реферативного журнала (около 200 классов).
 3. Глубокое индексирование по Рубрикатору ВИНТИ (более 50 тыс. классов), первые 3 уровня соответствуют ГРНТИ.
- * Из более чем 8 тыс. рубрик ГРНТИ около 3 тыс. рубрик 3-го уровня и около 450 рубрик 2-го уровня находятся в сфере интересов ВИНТИ.

История разработки

- **наборы данных для обучения:** метаданные (заглавия, рефераты или аннотации, ключевые слова), снабжённые классификационными кодами (отраслевых отделов, выпусков РЖ, рубрик ГРНТИ) – несколько миллионов документов.
- **2017-2019 гг.:** разработка и внедрение программы автоматической классификации.
 - основные инструменты: язык *Python*, эмбединг *Word2Vec*, библиотека *scikit-learn*;
 - программа классификации запускается в трёх режимах: диалоговый (*GUI*), пакетный (утилита командной строки) и сетевой сервис.
- **2024-2025 гг.:** переход с векторных моделей слов *Word2Vec* на языковые модели *BERT*:
 - *DeepPavlov/rubert-base-cased* для русского языка;
 - *google-bert/bert-base-uncased* для английского языка;
 - дообучение предобученных моделей методом *fine-tuning LoRA*.
 - основные инструменты: язык *Python*, библиотека *PyTorch*.
 - программа классификации запускается в пакетном режиме.

Функциональные компоненты ПО

❑ **Предобработчик** очищает и унифицирует текст:

- ❖ очистка от элементов разметки;
- ❖ перекодирование служебных символов, диакритических и математических знаков;
- ❖ лемматизация слов (приведение к словарной форме).

❑ **Векторизатор** преобразует тексты в числовые представления (эмбеддинги)

- ❖ **Word2Vec** – контекстно-независимые эмбеддинги;
- ❖ **BERT** – контекстно-зависимые эмбеддинги.

❑ **Классификатор** выполняет мультиклассовую классификацию текста, результат – список классов с их вероятностями.

❑ **Процедуры оценки качества** сравнивают результаты автоматической классификации с результатами ручной классификации и вычисляют **точность (precision)**, **полноту (recall)** и **F1-меру (f1 score)** для каждого класса в отдельности, а также средние значения для всех классов (макро и микро средние).

Запуск автоматической классификации

С 2019 года утилита командной строки запускается по расписанию через программную оболочку «Электронный эксперт», которая:

- отбирает из БД и формирует входной поток текстов (*заглавие + аннотация + ключевые слова*);
- запускает классификацию;
- загружает результаты классификации в БД.

Результат классификации - строка в формате

Код1-Вероятность1\Код2-Вероятность2\Код3-Вероятность3\...

Пример результата классификации с кодами ГРНТИ 2-го уровня:

68.35-0.33275\34.23-0.29211\34.31-0.12737

Пользователь увидит расшифровку:

Код ГРНТИ	Название рубрики	Вес/Релевантность/Вероятность
68.35	Растениеводство	0.33275
34.23	Генетика	0.29211
34.31	Физиология растений	0.12737

Использование результатов: подсказки для разметчика

Поиск F7

По УНМ 75120481

По СИД J22892114

ИД

Статей 11 Р11 ЭФ 11

БО выпуска PDF ОГЛ

// Anal. Lett. [Электронный ресурс]. — 2025. — 58, № 8. —

Завершение работы

Дата разметки 14.04.2025

ФИО sa — Администрат

Группа РМА— Автоматическая разм

Показывать Оглавление

F9 Стандартно Разметка

Содержание 2/11 Ctrl+F3 - Статистика по изданию Интернет

ИД статьи	Название	Разм.	Страницы
J2289211419	Use of Constant-Analyte Samples in Updating Quantitative Near-Infrared Calibration Models for	X	1260-1274
J2289211427	Manganese-Doped Carbon Quantum Dots (CQDs) Prepared from Bovine Bone Powder and Their	X	1275-1287
J2289211435	Adenosine-5'-Triphosphat		
J2289211443	Solid-Phase Microwave Synth		
J2289211451	Determination of Inorganic A		
J228921146X	Synchronous Fluorescence Sp		

Эл. разметка Статусы

Manganese-Doped Carbon Quantum Dots (CQD: Determination of Chlorpyrifos / Lv Jing. — с. 1275-1287. — Англ.; Abstract Employing a one-step pyrothermal method with bovine bone powder as the carbon source, Mn-doped fluorescent carbon quantum dots were synthesized. Experimental investigations were conducted on

Автоматическая разметка для J2289211427

Змечено в ОНИ : X => РЖ 19Г напр. реф. 170

Результаты автоматической разметки

ОНИ	Вес
X	0.654
БП	0.271

Всего : 2

Обновить Выход

Автоматическая разметка для J2289211427

Размечено в ОНИ : X => Р

Результаты автоматической разметки

ИД	РЖ	Вес
J2289211427	19ГД	0.635
J2289211427	84	0.227
J2289211427	18Л	0.089
J2289211427	19Б2	0.019

Всего : 4

Обновить Выход

Здесь результаты классификации по ОНИ (отделам научной информации) и выпускам РЖ (реферативного журнала) используются в качестве "подсказок" для разметчика.

Автоклассификация новых поступлений. Пример: рубрика 28.23 «Искусственный интеллект».

для рубрики 28.23

"Искусственный интеллект"

☐ Связать с текущей рубрикой

ID	БО
B2316605621	ЭВОЛЮЦИОННЫЙ АЛГОРИТМ МНОГОЭКСТРЕМАЛЬНОЙ МНОГОПАРАМЕТРИЧЕСКОЙ ОПТИМИЗАЦИИ / Репин А. И., Безуглов Е. А. // Автоматизированные системы управления технологическими процессами. Control-2025: Сборник материалов 5 Международной научно-практической конференции, Москва, 18 марта, 2025. - М.: МЭИ, 2025. - 76 с.
J22906565128	Метод обнаружения аномалий в метеорологических данных на основе нейронных сетей / Чжу Ч., Гуан И., Конг Л., Нань Ю. ... // Метеорология и гидрология. - Москва: Метеорол. и гидрол., 2025, N 2
J22906565144	Методология суррогатного моделирования нелинейной динамики атмосферы: от концептуальной модели к нейронным сетям / Солдатенко С. А., Ангудович Я. И. // Метеорология и гидрология. - Москва: Метеорол. и гидрол., 2025, N 2
J2308359749	УЛУЧШЕНИЕ КАЧЕСТВА РАСПОЗНАВАНИЯ ПАРАМЕТРОВ ФЛОТОМАШИНЫ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННОЙ СЕТИ / Затонский А. В., Кузнецов С. В., Саломатова К. А. // Горный информационно-аналитический бюллетень. - Москва: Моск. гос. горн. ун-т, 2025, N 6
J2309865915	Визуализация опухоли на основе обработки вейвлет-преобразования эластографического сигнала с помощью нейросети / Кравчук Д. А. // Прикладная физика. - Москва: ВИМИ, 2025, N 3
J23099019107	Сегментация сельскохозяйственных изображений на основе глубокого обучения / Гапон Н. В., Пузеренко А. В., Жданова М. М., Рудой Д. В. ... // Технологии живых систем. - Москва: Радиотехника, 2025, Том 22, N 2
J2310218456	Определение типа сканированного документа методом динамической трансформации временных осей / Максимова Т. Р., Безматерных П. В. // Труды Института системного анализа РАН, 2025, Том 75, N 2
J2311791281	Модель изображений бесструктурных областей для анализа пигментных узоров с применением искусственного интеллекта в онкодерматологии / Никитаев В. Г., Сергеев В. Ю., Кегелик Н. А., Козлов В. С. // Медицинская техника, 2025, N 2
J2311994X43	Результаты моделирования динамики роботов в средах Matlab и SolidWorks на примере подводного робота / Муса З., Маева А. // Научные технологии. - Москва: Радиотехника, 2025, Том 26, N 3
J2312578891	АТАКУЕМЫЕ СРЕДСТВА МАШИННОГО ОБУЧЕНИЯ: АВТОМАТИЗАЦИЯ РИСК-АНАЛИЗА ПРОЦЕССОВ РЕАЛИЗАЦИИ СЦЕНАРИЕВ / Нархов Д. А., Кульшин Д. В., Козина К. В., Неменуций М. Д. // Информация и безопасность, 2025, Том 28, N 2
J23128213107	Оценка снизу регрета алгоритма агрегирования экспертных прогнозов для переменного числа активных экспертов / Зухба Р. Д., Зухба А. В. // Труды Московского физико-технического института (государственного университета) (МФТИ), 2025, Том 17, N 2
J23128213123	Метод выявления аномальных движений людей на видеозаписях без предварительного обучения / Воробьев Н. Д. // Труды Московского физико-технического института (государственного университета) (МФТИ), 2025, Том 17, N 2

Примерно 40% входного потока ВИНТИ РАН не отражается в РЖ и БД ВИНТИ, но его можно автоклассифицировать и показать пользователям библиографию.

Рекомендации для научных сотрудников, выполняющих аналитико-синтетическую переработку документов

индексатор-1 B19461782261 (B19461782261)

Реферат: Представлен метод, позволяющий организовать движение беспилотных летательных аппаратов в строю. В основе предложенного решения лежит модификация метода потенциального поля. Также представлены результаты моделирования гетерогенного роя квадрокоптеров, использующего предложенный метод. В качестве сети информационного взаимодействия агентов используется Wi-Fi сеть

Загл. осн.: Формирование строя группой беспилотных летательных аппаратов

Класс-ция

Классификатор	Результат класс.	Релевантность
ОНИ ВИНТИ	Математика	0.51664
ОНИ ВИНТИ	Автоматика и радиоэлектроника	0.21912
ОНИ ВИНТИ	Машиностроение	0.14734
Реферативный журнал ВИНТИ	13Д (РЖ закрыт)	0.43976
Реферативный журнал ВИНТИ	81 Техническая кибернетика	0.29633
Реферативный журнал ВИНТИ	37 Робототехника	0.14979
ГРНТИ-3	50.05.15 Теория и проблематика г	0.3066
ГРНТИ-3	55.30.05 Констркции и техническ	0.18137
ГРНТИ-3	28.23.15 Распознавание образов.	0.1554
ГРНТИ-3	28.23.27 Интеллектуальные робот	0.05216
ГРНТИ-3	49.43.29 Системы радиосвязи	0.04923
ГРНТИ-3	29.37.27 Акустика природных сред	0.03824
ГРНТИ-3	50.49.37 Автоматизированные сис	0.03131
13Ж Математическая кибернети	28.23.20 Формирование решений	0.565856 (corr)
81 Техническая кибернетика	28.23.20 Формирование решений	0.565856 (corr)
13Ж Математическая кибернети	28.23.15 Распознавание образов.	0.334235 (corr)
81 Техническая кибернетика	28.23.15 Распознавание образов.	0.334235 (corr)
13Ж Математическая кибернети	28.23.27 Интеллектуальные робот	0.19364 (corr)
81 Техническая кибернетика	28.23.27 Интеллектуальные робот	0.19364 (corr)

Кол-во документов: 1 kw_count: 3235 - 0,00 сек.

Тематическая классификация текстов <http://solar.viniti.ru/atc/>

Текст для анализа:

Synthesis, mechanical, and thermophysical properties of high-entropy (Zr,Ti,Nb,Ta,Hf)C_{0.8} ceramic.
Two high-entropy carbides, including stoichiometric (Zr,Ti,Nb,Ta,Hf)C and nonstoichiometric (Zr,Ti,Nb,Ta,Hf)C_{0.8}, were prepared from monocarbides and ZrH₂. Their sinterability, microstructures, mechanical properties, thermophysical properties, and oxidation behaviors were systematically compared. With the introduction of carbon vacancy, the sintering temperature was lowered up to 300°C, Vickers hardness was almost unaffected, whereas the strength decreased significantly generally due to the decrease of covalent bonds. The thermal conductivity shows a 50% decrease for nonstoichiometry high-entropy carbide, which is a major consequence of the lower electrical conductivity. The oxidation resistance in high temperature water vapor was not sensitive to carbon stoichiometry.

Как пользоваться ?

Язык текста

Английский ▾

Классифицировать по схеме ...

☐ ... Область науки и техники ?

☐ ... ГРНТИ-1 ?

☒ ... ГРНТИ-2 ?

☒ ... ГРНТИ-3 ?

☐ ... Выпуски РЖ ВИНТИ ?

Порог выдачи результатов ?

0,1 ▴ ▾

Выполнить

Обработано за 0 сек.

Вероятность

Название

ГРНТИ-2:

0,26038

[Технология производства силикатных и тугоплавких неметаллических материалов \(шифр 61.35\)](#)

0,25026

[Физика твердых тел \(шифр 29.19\)](#)

0,14735

[Коррозия и защита от коррозии \(шифр 81.33\)](#)

0,10532

[Физическая химия \(шифр 31.15\)](#)

Все результаты

ГРНТИ-3:

0,91878

[Огнеупоры \(шифр 61.35.35\)](#)

0,50234

[Керамика \(шифр 61.35.29\)](#)

0,17772

[Неметаллические коррозионностойкие материалы \(шифр 81.33.43\)](#)

Все результаты

Мониторинг качества классификации

Если для публикаций по биологии, экономике промышленности, математике, физике, химии система даёт достаточно точные ответы, то для других областей (автоматика и радиоэлектроника, машиностроение, транспорт и др.) результаты получаются более размытыми, без явных приоритетов.

Это можно объяснить:

- **Несовершенством используемых алгоритмов и моделей** автоматической классификации;
- **Свойствами текстов.** Например, публикации по биологии или химии содержат существенно больше избирательных (селективных) терминов, чем публикации по машиностроению или транспорту, где лексическая многозначность терминов встречается чаще;
- **Недостатками классификационных схем**, связанными с условностями деления на отрасли в политематических и междисциплинарных областях (некоторые классы в значительной степени пересекаются).

Сравнение моделей классификации: BERT против Word2Vec

Проблема: моделям *BERT* для приемлемого качества классификации требуется гораздо больше документов по каждому классу в обучающем датасете. Поэтому пришлось увеличить порог до 300 экземпляров класса, что дало лишь 246 обученных классов ГРНТИ 2-го уровня.

Для комбинации *Word2Vec+Perceptron* был установлен порог 25 экземпляров, что дало 402 обученных класса.

Сравнение результатов классификации на случайной тестовой выборке из БД ВИНТИ объемом 36,66 тыс. русскоязычных текстов вида «название + аннотация + ключевые слова».

Метрика	Модель	модель Word2Vec и Perceptron с одним скрытым слоем	модель BERT и файнтюнинг LoRA
микро F1-мера		0,305	0,422
макро F1-мера		0,261	0,145
кол-во классов ГРНТИ-2, хотя бы раз присвоенных тестовым текстам		394	212
минимальное кол-во экземпляров класса в обучающем датасете		25	300
кол-во обученных классов ГРНТИ-2		402	246

Заключение

- ❑ Языковая модель **BERT**, более требовательная к минимальному количеству экземпляров класса в обучающей выборке, не была обучена классифицировать по низкочастотным классам. Поэтому на низкочастотных классах старая модель классификации (**Perceptron + Word2Vec**) показала преимущество по сравнению с более современной моделью **BERT**. Зато модель **BERT** показала существенное преимущество на высокочастотных классах, что проявилось в значительном повышении микро F1 меры с 0,305 до 0,422 (*при классификации рубриками ГРНТИ 2-го уровня*).
- ❑ Уровень доверия к результатам автоклассификации зависит как от субъективных предпочтений оценивающего, так и от объективных факторов, в первую очередь от степени совершенства алгоритмов и моделей классификации, качества и объёма обучающих наборов данных, глубины классификации и количества классов.
- ❑ В дальнейшем предполагается продвинуться в сторону автоматической разметки, когда некоторые материалы входного потока, выделенные по формальным признакам, будут направляться в тематические отделы без участия человека (*возможные признаки: названия журналов, издательств, определённые тематики по кодам УДК и т.п.*)
- ❑ Полный отказ от ручной классификации вряд ли возможен в ближайшее время, поэтому при формировании информационных продуктов и услуг необходимо находить адекватные механизмы сосуществования и взаимного дополнения людей и программ-роботов, выполняющих классификацию.



Доклад окончен.
Спасибо за внимание!

Список литературы

1. Егоров В.С., Козлова Е.С., Ломотин К.Е., Федорец О.В., Филимонов А.В., Шапкин А.В. Система автоматической классификации текстов для обработки потока научных публикаций в ВИНТИ РАН // Научно-техническая информация. Сер. 2. Информационные процессы и системы. 2020. № 5. С. 1-12. DOI: 10.36535/0548-0027-2020-05-1
2. Кусакин И.К., Федорец О.В., Романов А.Ю. О классификации коротких научных текстов // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. 2023. № 7. С. 22-28. DOI: 10.36535/0548-0019-2023-07-3