

УДК 004.931

Программная обработка полнотекстовых научных публикаций для формирования тематических баз данных свойств материалов

Хвостова Мария Олеговна^(*)

mhvastova@kaf65.ru

SPIN-код: 1509-8350

Абрамов Антон Владиславович

xxxzx@mail.ru

SPIN-код: 5438-2728

Артамонов Алексей Анатольевич

aaartamonov@mephi.ru

SPIN-код: 9375-6151

Сучков Алексей Николаевич

ANSuchkov@mephi.ru

SPIN-код: 8012-3494

НИЯУ МИФИ, Москва, Россия

Аннотация. Представлен метод автоматизированного выделения свойств материалов из полнотекстовых научных статей, основанный на разработанных алгоритмах распознавания, стандартизации и связывания элементных составов материалов с их физическими характеристиками. Рассмотрено программное средство извлечения выбранных свойств сплавов-припоев и спаянных соединений из текста и таблиц, реализующее данный метод. Предложены последовательность обработки и способ токенизации текста, позволившие достичь полноты и точности извлечения целевой информации > 90 % на тестовой выборке. Показано, что представленный метод обработки может быть применен к публикациям, относящимся к иным областям материаловедения. Сделан вывод о возможности формирования тематических баз данных свойств материалов с использованием метода.

Ключевые слова: цифровое материаловедение, распознавание образов, обработка данных

Automated Processing of Full-Text Scientific Articles for Materials Databases Population

Khvastova Mariia Olegovna^(*)

mhvastova@kaf65.ru

SPIN-code: 1509-8350

Abramov Anton Vladislavovich

xxxzx@mail.ru

SPIN-code: 5438-2728

Artamonov Alexey Anatolyevich

aaartamonov@mephi.ru

SPIN-code: 9375-6151

Suchkov Alexey Nikolaevich

ANSuchkov@mephi.ru

SPIN-code: 8012-3494

MEPhI, Moscow, Russia

Abstract. The paper proposes a method for the automated extraction of materials properties from full-text scientific articles based on elemental composition and quantitative data recognition, standardization, and linking algorithms developed by the authors. A software

tool for the extraction of selected solders' and soldered joints' properties from text and tables built upon the method in presented. A tokenization technique allowing for enhanced precision and recall on test data ($> 90\%$) is discussed. The applicability of the method proposed for the processing of publications pertaining to other material science fields is demonstrated. The authors conclude on the possibility of populating materials databases using the method.

Keywords: digital material science, pattern recognition, data processing

Введение. Автоматизация обработки научно-технических источников облегчает получение количественных характеристик материалов, необходимых для проведения расчетов и аналитики предметной области [1]. Распознавание, стандартизация и связывание физических свойств и элементных составов являются ключевыми задачами при реализации программного инструмента для создания структурированных наборов данных о материалах на основе информации из научных публикаций. В настоящее время описанная проблематика активно исследуется в работах, посвященных металлическим сплавам [2–4], сверхпроводящим и другим материалам [5–7]. Отличие предложенного метода от изученных состоит в алгоритмической обработке текстовой и табличной информации из каждой статьи, объединении результатов по материалу и применении механизмов конвертации в атомные/массовые проценты для элементных составов сплавов и в Международную систему (СИ) для физических характеристик.

Материалы и методы. Выполнены две программные реализации предложенного метода. Входными данными для обработки являлись полнотекстовые научные публикации в PDF на английском языке, найденные по запросу к Scopus API и выгруженные из портала archive.org автоматически. Элементные составы и физические величины выделялись из текста и таблиц статей с помощью регулярных выражений, составленных вручную. Восстановление физических величин до физических характеристик и связывание материалов и характеристик осуществлялось: А) на основе относительных расстояний и статистики объектов в предложении; Б) на основе расположения в иерархии токенов и статистики объектов в предложении. Иерархическая токенизация подразумевала разбиение каждого предложения по служебным частям речи, которые лексически выражают отношения между его частями, а частей предложения — на именные группы и остальные слова (рис. 1).

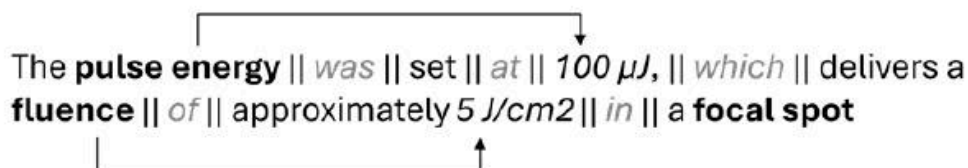


Рис. 1. Пример токенизации предложения с выделением **именных групп**

Возникавшая в варианте Б множественность выбора при восстановлении физических характеристик разрешалась с помощью оценки косинусных рас-

стояний между характерными для единицы измерения величинами и возможными параметрами. В реализации А выделение и связывание сущностей осуществлялись на уровне предложения в рамках одной итерации. В реализации Б выделение сущностей выполнялось на отдельной итерации по всем предложениям текста, предшествующей связыванию.

При создании программного инструмента, реализующего предложенный метод, применялись языки программирования Python и JavaScript, библиотеки regex, NumPy, OpenCV, PaddleOCR, Stanza и др., модель Word2Vec, СУБД MongoDB, среда выполнения Node.js, фреймворки Express.js и FastAPI, инструмент фронтенд-разработки React.

Результаты. Апробация метода проведена на 15 научных публикациях, посвященных пайке. Реализация А показала полноту извлечения около 90 % для всех целевых видов информации, однако точность на объектах «материал-свойство» оказалась недостаточной, не превысив 75 % (табл. 1). Для реализации Б точность достигла 90 % при сохранении полноты (табл. 2). Перечень оцениваемых характеристик был закрытым: температуры солидус и ликвидус припоев, время и температура пайки и полученная прочность спаянного соединения.

Таблица 1

Реализация А: Оценка качества извлечения информации
(AV — среднее значение, R — полнота, P — точность)

Тип AV	Всего	
	R	P
Припой	0,97	0,86
Спаянное соединение	0,96	0,83
Фаза, Продукт реакции	0,94	0,94
Физическая характеристика	0,96	0,99
Припой + характеристики	0,9	0,63
Спаянное соединение + характеристики	0,97	0,74

Таблица 2

Реализация Б: Оценка качества извлечения информации
(AV — среднее значение, R — полнота, P — точность)

Тип AV	Всего	
	R	P
Припой	1	0,9
Спаянное соединение	1	1
Припой + характеристики	1	0,94
Спаянное соединение + характеристики	1	0,93

Обсуждение. Изменения значений метрик, полученные в реализации Б, объясняются более корректным использованием контекста естественного языка. Средние значения по рассмотренным объектам превышают аналогичные, достигнутые в работах Liu et al. и Wang et al., посвященных сплавам, с помощью обученных моделей разных архитектур [2, 3]. В отличие от изученных методов, иерархическая токенизация позволяет распознавать заранее не определенные свойства материалов. Описание сплавов с помощью регулярного выражения обеспечивает универсальность распознавания данного класса материалов, тогда как изученные методы требуют переобучения моделей для конкретных составов. Подобный подход масштабируем на материалы других классов, символьные записи которых представляют последовательность химических элементов периодической таблицы и числовых коэффициентов или внесены в базу данных распознаваемых материалов вручную. Введенная классификация материалов (сплав/керамика/стекло/композит) упрощает фильтрацию результатов обработки (рис. 2).

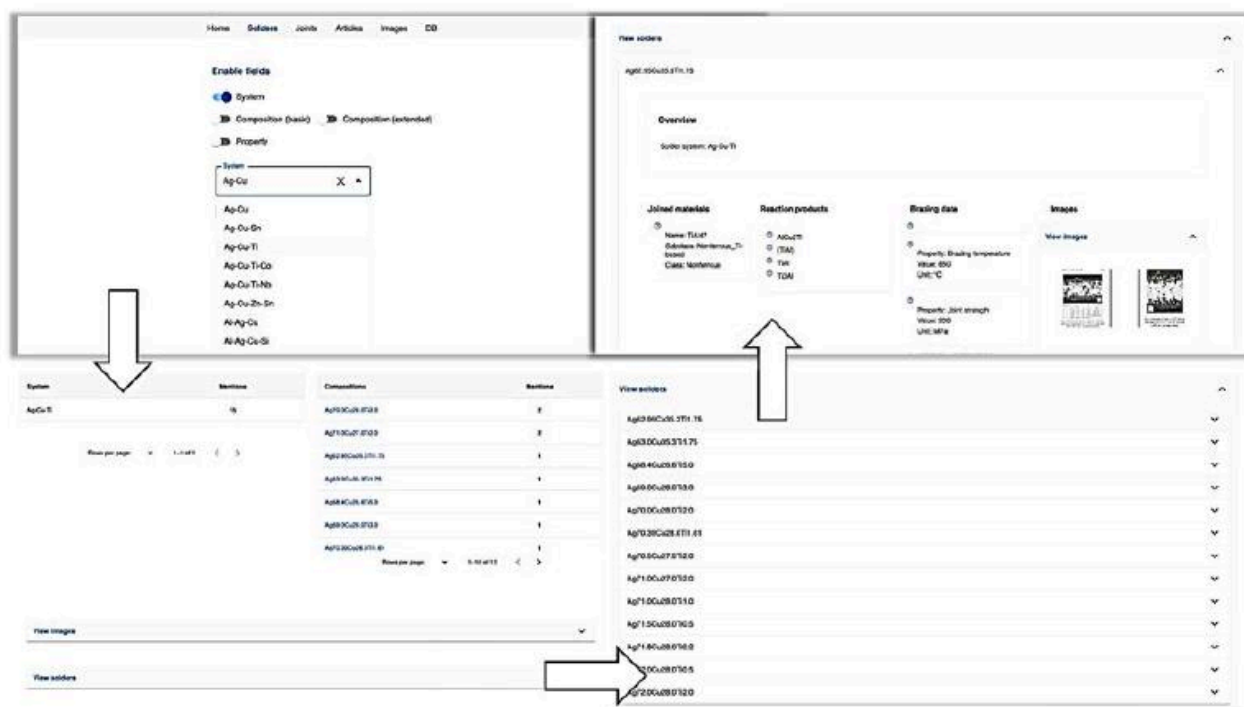


Рис. 2. Пример фильтрации результатов обработки публикаций, посвященных пайке

Заключение. Результаты, полученные на тестовой выборке для реализации предложенного метода, обладающих разной степенью контекстной осведомленности, были ожидаемы и свидетельствуют о возможности использования реализации Б для формирования тематической базы данных о припоях. Авторы видят необходимость провести тестирование метода в условиях расширенных выборки, целевых свойств и материалов, чтобы оценить потенци-

ал обработки научных публикаций, описывающих эксперименты не только со сплавами, но и органическими соединениями, ядерными материалами и другими, с его помощью.

Список источников

- [1] Хвостова М.О., Абрамов А.В., Артамонов А.А., Сучков А.Н. Цифровая платформа консолидации знаний о припоях для решения прикладных и исследовательских задач современного материаловедения. *Русский инженер. II Всерос. конгресс с междунар. участием: тез. докл.* Москва, Изд-во МГТУ им. Н.Э. Баумана, 2024, 207 с.
- [2] Liu Y., Wen S., Yin J. et al. ASaRE-Net: automatic information extraction from Al-Si alloy materials science literature for corpus construction. *Journal of Materials Science*, 2024, vol. 59, pp. 15330–15346. <https://doi.org/10.1007/s10853-024-10074-z>
- [3] Wang W., Jiang X., Tian S. et al. Automated pipeline for superalloy data by text mining. *npj Computational Materials*, 2022, vol. 8. <https://doi.org/10.1038/s41524-021-00687-2>
- [4] Shah S., Vora D., Gautham B.P. et al. A Relation Aware Search Engine for Materials Science. *Integrating Materials and Manufacturing Innovation*, 2018, vol. 7, pp. 1–11. <https://doi.org/10.1007/s40192-017-0105-4>
- [5] Foppiano L., Castro P. B., Ortiz Suarez P., Terashima K., Takano Y. Ishii M. Automatic extraction of materials and properties from superconductors scientific literature. *Science and Technology of Advanced Materials: Methods*, 2023, vol. 3. <https://doi.org/10.1080/27660400.2022.2153633>
- [6] Court C.J., Cole J.M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data*, 2018, vol. 5, pp. 1–11. <https://www.nature.com/articles/sdata2018111#Sec13>
- [7] Huang D., Cole J.M. A database of thermally activated delayed fluorescent molecules auto-generated from scientific literature with ChemDataExtractor. *Scientific Data*, 2024, vol. 11. <https://doi.org/10.1038/s41597-023-02897-3>