

© *Веселов Е.М.¹, Грибова В.В.², Тарасов А.В.³, Тимченко В.А.⁴*

ФОРМАЛИЗАЦИЯ ПРОТОКОЛОВ ТЕХНОЛОГИЧЕСКИХ ОПЕРАЦИЙ ЛАЗЕРНОГО АДДИТИВНОГО ПРОИЗВОДСТВА С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

¹ Институт автоматизации и процессов управления ДВО РАН, evgeniyvsl@vk.com;

² Институт автоматизации и процессов управления ДВО РАН, gribova@iacp.dvo.ru, SPIN-код: 7481-5878;

³ Институт автоматизации и процессов управления ДВО РАН, atarasov@iacp.dvo.ru;

⁴ Институт автоматизации и процессов управления ДВО РАН, vadim@iacp.dvo.ru, SPIN-код: 5266-9228.

Предложен метод для перевода в машинопонимаемое представление текстовых документов, содержащих описания технических заданий и отчетов о технологических операциях лазерной аддитивной обработки металлических материалов. Ключевая особенность метода состоит в сочетании показавшего свою эффективность двухуровневого онтологического подхода к формированию формализованного машинопонимаемого представления баз данных по лазерному аддитивному производству с возможностями больших языковых моделей по обработке и анализу текстов. Метод реализован в виде распределенного композитного облачного сервиса.

Ключевые слова: большие языковые модели, векторные базы данных, анализ текстов, извлечение данных, структуризация информации, лазерная аддитивная обработка материалов.

Введение. Развитие технологий лазерной аддитивной обработки металлических материалов сопровождается необходимостью выявления пригодных на практике режимов работы технологического оборудования в условиях выполнения сложных и часто уникальных производственных задач. От корректности выбора параметров, определяющих режим выполнения технологического процесса, напрямую зависит качество и воспроизводимость получаемого результата. На практике инженер-технолог опирается на собственный опыт и знания, техническую документацию, различные руководящие документы, а также государственные и отраслевые стандарты. Однако ввиду большого разнообразия материалов, конфигураций изделий и условий проведения работ выявление надлежащих режимов (обеспечивающих соответствие результата требованиям технического задания) представляет собой интеллектуально сложную задачу и является серьезным барьером на пути массового внедрения лазерных аддитивных технологий в производственные процессы [1–3].

Вместе с тем, в целом, можно констатировать, что имеет место острая нехватка гибко адаптируемых (масштабируемых) интеллектуальных ассистентов, способных оказывать помощь операторам оборудования, технологам и другим отраслевым специалистам в расчете пригодных для конкретной практической задачи технологических параметров (режимов обработки) [4, 5]. В связи с этим, при участии коллектива авторов ведется разработка жизнеспособной интеллектуальной системы поддержки принятия решений (СППР) [6], дающей рекомендации по настройке надлежащих режимов работы технологического оборудования, реализующего процессы лазерной аддитивной обработки металлических деталей (с использованием являющейся в настоящее время одной из наиболее перспективных технологий DED-LB – laser-based directed energy deposition, “процесс прямой подачи энергии лазера и материала” [7]).

Одним из ключевых компонентов создаваемой СППР является база прецедентов, которая представляет собой централизованную облачную библиотеку протоколов технологических операций (ТО), в которой хранятся структурированные машинопонимаемые описания подробных отчетов о проведенных ТО обработки материалов (лазерной наплавки, сварки). Поскольку рассматриваемая база данных лежит в основе методов комплексного интеллектуального анализа данных, применяемых в СППР – в частности, гибридного метода извлечения прецедентов – её объём оказывает непосредственное влияние на эффективность работы системы. Кроме того, при достижении достаточного объёма базы, она может быть

использована в качестве датасета для интеграции в комплекс программных средств, основанных на методах машинного обучения. Однако в настоящее время пополнение базы данных осуществляется экспертами вручную с использованием структурного редактора, что сопряжено с рядом ограничений. Во-первых, данный процесс требует значительных временных затрат. Во-вторых, в ряде случаев экспертам необходимо разбираться в структуре нормативно-справочных баз данных (БД), содержащих сведения об оборудовании и материалах для аддитивного производства. Это создаёт избыточную когнитивную нагрузку на специалистов, приводит к существенным затруднениям в работе и, как следствие, снижает частоту и своевременность пополнения базы протоколов.

В то же время существуют архивы текстовых документов, включающих протоколы (технические отчёты) проведённых ТО, а также технические задания на выполнение ТО, формируемые различными организациями (впоследствии включаемые в состав отчётов). Эти данные требуют преобразования в формализованное, машинопонимаемое представление для их эффективного использования.

Таким образом, разработка программного обеспечения для автоматизации процесса извлечения релевантной информации из технической документации (отчетов) о проведенных ТО, их преобразования в машинопонимаемое структурированное представление для последующего пополнения базы прецедентов СППР является значимой и актуальной научно-практической задачей, решению которой посвящено настоящее исследование.

1. Материалы и методы. Общая схема наполнения библиотеки протоколов ТО лазерной аддитивной обработки посредством извлечения релевантной информации из текстовых протоколов ТО и формирования их структурированного машинопонимаемого представления представлена на рисунке 1.

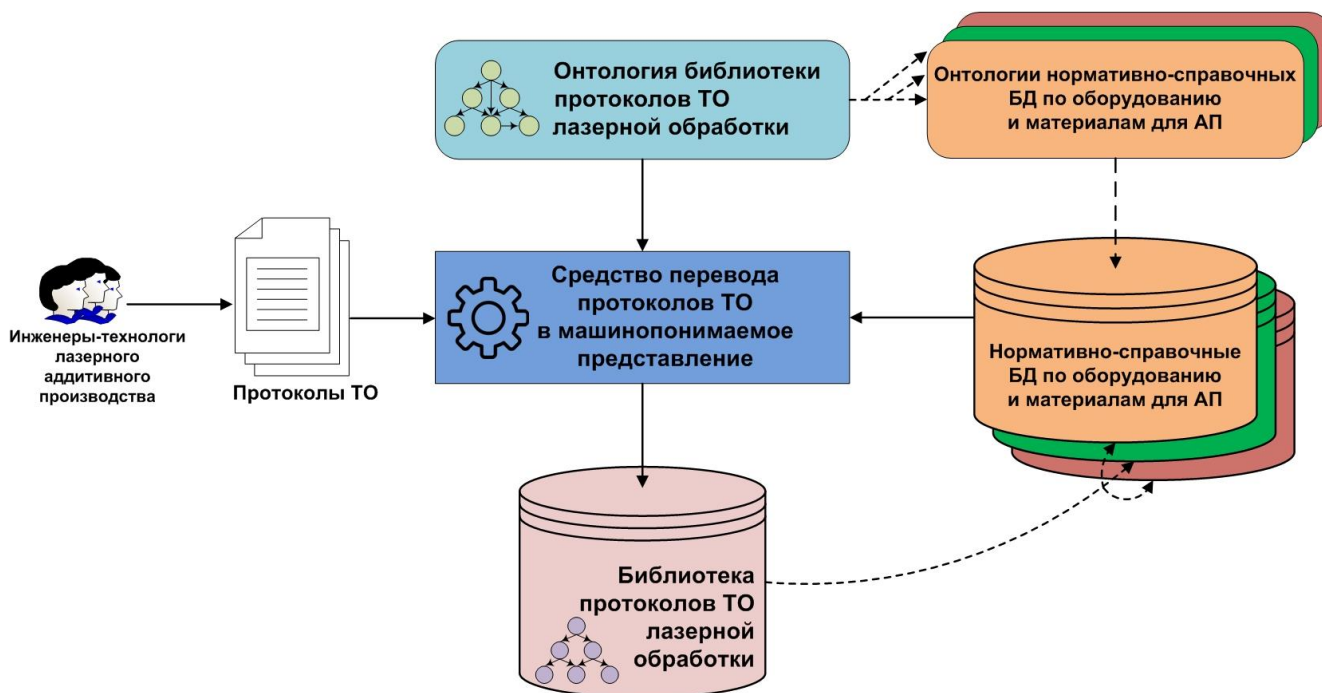


Рис. 1 – Схема наполнения библиотеки протоколов ТО лазерной аддитивной обработки с использованием средства их перевода в машинопонимаемое представление

Протокол ТО представляет собой технический документ (отчет), содержащий текстовое описание всех существенных сведений о выполненной ТО. Помимо общих фактов о ТО (название, номер протокола, срок, цель и место выполнения ТО) протокол, в общем случае, включает в себя разделы с описанием условий окружающей среды при выполнении ТО, технического задания на выполнение ТО, использованных средств и параметров выполнения ТО (режима работы лазерного робототехнического оборудования), а также результатов выполнения ТО (с их детальной спецификацией и оценкой). Разработка структурной онтологической модели (онтологии) библиотеки протоколов ТО лазерной обработки (и, в том

числе, формализованного протокола ТО) выполнена в соответствии с общими рекомендациями по проектированию и изготовлению деталей для процесса “прямой подачи энергии и материала” (ГОСТ Р 70242-2022). Её подробное описание, а также описание нормативно-справочных БД по оборудованию и материалам для АП вместе с их онтологическими моделями представлено в работе [6]. В контексте применяемого двухуровневого подхода онтологическая модель задаёт формальное описание структуры и ограничений на возможное содержание формируемой на её основе целевой/предметной БД (библиотеки), а также определяет правила формирования и интерпретации данных. Таким образом, онтологическая модель определяет формализованное машинопонимаемое представление содержимого целевых БД в виде семантических орграфов понятий.

При разработке средства перевода протоколов ТО в машинопонимаемое представление было проведено сравнение двух основных подходов: основанного на правилах (rule-based) и основанного на больших языковых моделях (Large Language Model, LLM-based). В результате сравнения для реализации был выбран подход на основе LLM, как более гибкий и перспективный для анализа и извлечения информации из неструктурированных текстов на естественном языке [8, 9]. Для упорядоченного хранения и администрирования документов, представленных в векторизованном формате (в форме т.н. эмбеддингов, “embeddings”), с целью их последующей передачи в запросах (prompt'ax) к LLM в качестве контекста, используется векторная БД Chroma. Для организации обмена (импорта и экспорта) данными в специализированном JSON-формате (ориентированном на представление семантических орграфов понятий) с облачной платформой IACaaS, на которой размещены все онтологические модели и целевые БД используется основанное на HTTP-запросах API удаленного взаимодействия с платформой [10]. Для отправки запросов к разным LLM, которые развернуты на выделенном вычислительном узле в ИАПУ ДВО РАН, также используется специально разработанное API на языке Python.

2. Результаты. Перед извлечением информации из текстовых документов (протоколов ТО) и формированием их структурированного машинопонимаемого представления в формате JSON [10] такие документы преобразовываются в векторное представление (embeddings) и помещаются в БД Chroma. На данном этапе используется библиотека LangChain. Предварительно тексты разбиваются на чанки длиной 800 символов (chunk_size=800) с пересечением между соседними чанками длиной в 80 символов (chunk_overlap=80). Для получения непосредственно векторов используется модель Nomic Embed Text v2, созданная специально для векторизации текстов.

Далее, на шаге 1 выполняется генерация JSON-шаблона, специально отмеченные места в котором необходимо будет заполнить извлеченной из текста протокола ТО информацией. JSON-шаблон формируется на основе Онтологии библиотеки протоколов ТО лазерной обработки, нормативно-справочных БД по оборудованию и материалам для АП и специализированного “служебного” информационного ресурса. В последнем явным образом специфицируются те места, в которых для корректного формирования протокола ТО в JSON-формате необходимо устанавливать ссылки на соответствующие фрагменты нормативно-справочных БД, а также, собственно, те места в шаблоне, которые должны быть заполнены извлеченными из текста протокола ТО данными. Отметим, что перечисленные выше ресурсы, используемые для генерации JSON-шаблона, используются также и для генерации списка вопросов, ответы на которые (в указанном документе) должна найти и поместить в соответствующие места LLM, к которой выполняется запрос.

На шаге 2 формируется шаблон с инструкциями (prompt) к LLM для поиска в указанном документе ответов на список вопросов. Шаблон имеет следующий вид:

```
PROMPT_TEMPLATE_QUESTIONS = """
**Контекст для генерации: **
{context}

**Список вопросов: **
{questions}
```

Инструкции по работе:

Заполни информацию о протоколах технологических операций, соблюдая следующие правила:

1. Выше был предоставлен список вопросов. Ответь на них в формате `"*Вопрос* > *Ответ*"`. Отвечай только по данной структуре. Больше ничего не меняй.

...

Остальные 9 правил, вместо которых стоит многоточие, опущены для краткости. В переменную `context` подставляется полученный из БД Chroma документ, а в переменную `questions` – сгенерированный ранее список вопросов. После этого перечислен набор правил, которых LLM должна придерживаться при генерации ответа. Результатом генерации является список пар вида `"*Вопрос* > *Ответ*"` (QA-пар), в которых формулировка вопроса строго соответствует формулировке из сгенерированного на предыдущем шаге списка вопросов. На результат работы LLM можно влиять, регулируя значения следующих гиперпараметров: размер контекстного окна, температура, `top_p` и штраф за повторение (`repeat_penalty`).

На шаге 3 выполняется обработка полученного списка QA-пар и заполнение на его основе отмеченных (специальным маркером) мест в сгенерированном на шаге 1 JSON-шаблоне. Также в те места, где должна быть установлена ссылка на фрагмент соответствующей нормативно-справочной базы, записывается предварительно построенный путь в этой базе до нужного понятия (если оно было найдено в этой базе).

Для оценки качества сгенерированных LLM ответов применяется перекрёстная проверка с использованием другой LLM. Проверка основана на анализе двух получаемых списков QA-пар.

Средство перевода протоколов ТО в машинопонимаемое представление реализовано как распределенный композитный сервис [10], точкой входа в который является web-приложение, разработанное с использованием фреймворка Flask (backend – на языке Python, frontend – HTML+CSS+JavaScript). В приложении реализованы функции полноценного администрирования локальной векторной БД, а также возможности по формированию протоколов ТО в специализированном машинопонимаемом JSON-формате (рисунок 2), которые можно (оставаясь в приложении), используя соответствующий метод API, экспортировать в библиотеку протоколов ТО лазерной обработки, размещенную на платформе IACPaaS.

Выберите документ

Поиск документов...

2. ТЗ 1_Деталь 11.8804.И.152.002

20150918 Наплавка порошка железа марки ПГ-ЮНХ16СР3 на металлическую подложку из нержавеющей стали марки 12Х18Н10Т. Э1-один слой

20150918 Наплавка порошка железа марки ПГ-ЮНХ16СР3 на металлическую подложку из нержавеющей стали марки 12Х18Н10Т. Э2-второй слой

2018.05.21 наплавка алюминиевого порошка АК 4-1 на алюминиевую подложку из сплава АК 4-1, Э4, один слой

2018.05.21 наплавка алюминиевого порошка АК 4-1 на алюминиевую подложку из сплава АК 4-1, Э5, один слой

2018.05.21 наплавка алюминиевого порошка АК 4-1 на алюминиевую подложку из сплава

Тип обрабатываемой операции:

Уровень 1:
Наплавка

Уровень 2:
Восстановительный ремонт изношенных и поврежденных деталей

Уровень 3:
-- Выберите --

Сгенерировать JSON

Список вопросов:

Место выполнения технологической операции

Срок выполнения технологической операции

Цель выполнения технологической операции

Температура окружающей среды

Относительная влажность

Атмосферное давление

Подложка:Толщина

Ширина

Глобальные настройки | **Настройки приложения**

Стратегия поиска:

☐ Поиск по списку вопросов ☒ Весь документ ?

Используемые мощности:

☐ Локальная модель: qwen2.5 ☒ Сервер ИАПУ ДВО РАН: llama_4_scout_17b_16e

Путь для сохранения:

JSON-answers/universal **Сохранить**

☐ Использовать валидацию результатов (модель: llama_3.3_70b)

Рис. 2 – Страница генерации протокола ТО в формате JSON

На вкладке «Настройки приложения» можно выбрать LLM, используемую для генерации QA-пар, из множества развернутых на вычислительном узле в ИАПУ ДВО РАН моделей, LLM, используемую для валидации, а также установить значения гиперпараметров, которые применяются ко всем моделям.

Заключение. В результате проведенного исследования разработан и реализован метод извлечения релевантной информации из технических заданий и отчетов о проведенных ТО лазерной аддитивной обработки металлических материалов и формирования её структурированного машинопонимаемого представления – в форме онтологических орграфов понятий. Предложенный метод основан на двухуровневом онтологическом подходе к представлению информационных ресурсов, содержащих формализованные отчеты (протоколы) о проведенных ТО, а также нормативно-справочную информацию об используемых в лазерном АП оборудовании и материалах. Для обработки и анализа технических текстов применяются большие языковые модели. В созданном программном решении реализована подсистема оценки качества результатов, основанная на их перекрёстной проверке с использованием валидационной модели. В настоящее время программное средство находится в стадии активного тестирования.

Помимо решения важной задачи по преобразованию в машинопонимаемое представление технических заданий для их дальнейшего использования в разрабатываемой СППР, практическая значимость работы состоит в возможности использования её результатов для автоматизации процесса пополнения одного из ключевых компонентов СППР – базы прецедентов.

Работа выполнена в рамках государственного задания ИАПУ ДВО РАН (тема № FWW-2025-0004).

Список литературы

1. Ethiraj N. A comprehensive review on application of machine intelligence in additive manufacturing / N. Ethiraj, T. Sivabalan, J. Sofia, D. Harika, M.P. Nikolova // Turkish Journal of Engineering. – 2025. – Vol. 9, no. 1. – P. 37–46.
2. Sibisi T.H. LAM additive manufacturing: A fundamental review on mechanical properties, common defects, dominant processing variables, and its applications / T.H. Sibisi, M.B. Shongwe, L.C. Tshabalala, I. Mathoho // The International Journal of Advanced Manufacturing Technology. – 2023. – Vol. 128. – P. 2847–2861.
3. Fan H.L. Analysis of powder flow from the nozzle to the substrate in multiangle laser cladding / H.L. Fan, Y.B. Liu, Z.W. Wang, R.H. Han, M.J. Zheng, H.B. Qi // Journal of Laser Applications. – 2024. – Vol. 36. – Art. 042080.
4. Soori M. Additive Manufacturing Modification by Artificial Intelligence, Machine Learning, and Deep Learning: A Review / M. Soori, F.K.G. Jough, R. Dastres, B. Arezoo // Additive Manufacturing Frontiers. – 2025. – Vol. 4, no. 2. – Art. 200198.
5. Chen K. A review of machine learning in additive manufacturing: design and process / K. Chen, P. Zhang, H. Yan, G. Chen, T. Sun, Q. Lu, Y. Chen, H. Shi // The International Journal of Advanced Manufacturing Technology. – 2024. – Vol. 135. – P. 1051–1087.
6. Gribova V. A Multi-Model Ontological System for Intelligent Assistance in Laser Additive Processes / V. Gribova, Y. Kulchin, A. Nikitin, P. Nikiforov, A. Basakin, E. Kudriashova, V. Timchenko, I. Zhevtun // Applied Sciences. – 2025. – Vol. 15, no. 8. – Art. 4396.
7. Ahn D.G. Directed Energy Deposition (DED) Process: State of the Art / D.G. Ahn // International Journal of Precision Engineering and Manufacturing-Green Technology. – 2021. – Vol. 8. – P. 703–742.
8. Dagdelen J. Structured information extraction from scientific text with large language models / J. Dagdelen, A. Dunn, S. Lee [et al.] // Nature Communications. – 2024. – Vol. 15. – Art. 1418.

9. Janatian S. From Text to Structure: Using Large Language Models to Support the Development of Legal Expert Systems / S. Janatian, H. Westermann, J. Tan, J. Savelka, K. Benyekhlef // Legal Knowledge and Information Systems. – 2023. – Vol. 379. – P. 167–176.

10. Москаленко Ф.М. Разработка композитных интеллектуальных сервисов, функционирующих в гетерогенной среде / Ф.М. Москаленко, В.А. Тимченко // Программная инженерия. – 2024. – Т. 15, № 8. – С. 411–424.