Гибридное извлечение состава вентиляционных установок из проектной документации: формальные грамматики и большие языковые модели

В. В. Осокин*, Р. Ф. Алимов‡

Заводы вентиляционного оборудования переносят данные об установках из проектов в программу подбора вручную. Описана схема автоматического извлечения этих данных: большая языковая модель читает изображение листа в JSON-схему и классифицирует листы, не опознанные по тексту; отбор остальных листов, разбор обозначений установок по формальным грамматикам, объединение источников и решения выполняет проверяемый код. Полнота отбора листов с таблицами систем и спецификациями — 100 % на обучающем корпусе и 94–99 % на отложенной выборке; при перепроверке 560 прочитанных полей ошибок чтения чисел не найдено.

Ключевые слова: большие языковые модели, извлечение информации, формальные грамматики, проектная документация, вентиляционные установки.

1. Введение

Завод вентиляционного оборудования получает проект раздела ОВ (отопление и вентиляция) в PDF и переносит сведения об установках в программу подбора вручную. Оформление у каждой организации свое, часть листов — сканы. Отдать документ большой языковой модели — получить невоспроизводимый ответ; разбирать его одними правилами — сломаться на первом непривычном формате. Модель решает только две задачи восприятия: читает изображение листа в заданную форму и относит к классам листы, которые не удалось опознать по тексту. Отбор остальных листов, разбор обозначений установок по формальным грамматикам, объединение источников и все решения выполняет проверяемый код.

* *Осокин Виктор Владимирович* — ООО «Два Облака», e-mail: victor@dvaoblaka.ru, ORCID: 0009-0000-6072-8237.

Osokin Viktor Vladimirovich — Dva Oblaka LLC.

‡ *Алимов Рустамжон Фахриддинович* — ООО «Два Облака», e-mail: rustam@dvaoblaka.ru, ORCID: 0009-0002-4941-8857.

Alimov Rustamzhon Fakhriddinovich — Dva Oblaka LLC.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

2. Постановка задачи

По проекту нужно получить список вентиляционных систем и для каждой — обозначение, состав установки как перечень секций по ходу воздуха и числовые параметры (расход, давление, температуры, мощности), по которым установка подбирается в каталоге завода.

Нужные сведения лежат на листах трех видов: таблица «Характеристика отопительно-вентиляционных систем» (ХОВС) по ГОСТ 21.602 [1], спецификация по ГОСТ 21.110 и вложенные техданные производителя.

Требования: не пропускать листы с системами (лишний лист инженер снимает одним щелчком, пропущенный теряет установки); не искажать числа и состав; работать на документах новых проектировщиков без донастройки. Если параметра в документе нет, правильный ответ — `null`, а не правдоподобное число, которое уйдет в подбор.

3. Метод

Шаг 1. Отбор листов. Из PDF берется текст; детекторы по заголовку таблицы ХОВС и признакам шапки (обозначения систем вида П1, единица $м^3/ч$) и по шапке спецификации («Поз.», «Наименование», «шт») дают класс листа. Листы с менее чем 400 буквами, нечитаемым слоем, крупным растром или обозначениями систем без заголовка классифицирует по изображению модель Claude Sonnet 4.6. Список подтверждает инженер.

Шаг 2. Чтение листов. Подтвержденные листы читает по изображению модель Claude Opus 5.5. Ответ ограничен JSON-схемой (ограниченное декодирование [2]): для ХОВС — обозначение, направление, расход ($м^3/ч$), давление (Па), мощность (кВт), нагреватели, рекуператор, охладитель, фильтры с температурами (°C) и мощностями; для спецификации — позиции с теми же величинами и перечнем секций 14 типов. Отсутствующее значение — `null`, выдумывать запрещено.

Шаг 3. Разбор обозначений по формальным грамматикам. У ряда производителей марка установки сама является формулой состава, например `ANR6 L/2K1/2P1/2F1/R1/N5.2/V1.0.P40.R-4x30/H1/B1 + P/...`: секции перечислены по ходу воздуха через «/», «+» отделяет вытяжку, числа задают размеры теплообменников. Для таких обозначений по каталогам производителей составлены четыре формальные грамматики: секционных установок (AIRNED, ANR, ANP), моноблочных (LITENED, UTR), компактных (MININED, MPU) и буквенных марок (VTS). Терми-

нальные символы грамматик — коды секций из конечного словаря (11 серий, 81 код: K1 клапан, F1 фильтр, N1 нагреватель, R1 рекуператор, V1 вентилятор и др.) и числовые поля фиксированного вида; нетерминалы — установка, приточная и вытяжная части, секция, типоразмер. Правила не содержат рекурсии, поэтому все четыре грамматики регулярны, и марка разбирается конечным автоматом без участия модели [3]. Кириллические двойники латинских букв нормализуются; цепочка, не выводимая ни в одной грамматике, дает предупреждение, а не отказ.

Шаг 4. Объединение источников. Расход, давление и мощность — из ХОВС; спецификация и техданные лишь заполняют пустые поля, расхождение больше 2 % попадает в конфликты. Состав: техданные производителя (если вложены), иначе марка, иначе секции спецификации, иначе блоки ХОВС. Температуры и мощности нагревателей и охладителей: спецификация, ХОВС, техданные; чего нет нигде — `null`.

4. Результаты

Отбор листов. Показатели — до подтверждения инженером, относительно ручной разметки. На обучающем корпусе (49 документов, около 2900 листов, из них модели показано 970) листы ХОВС и спецификаций найдены полностью: полнота 100 %, точность 87,7 и 96,5 %. На отложенной выборке (14 документов других проектировщиков, около 1400 листов) без изменений кода: полнота 94,4 и 99,1 %, точность 72,3 и 89,2 %.

Чтение листов. Из каждого из 64 документов взяты две системы, их листы прочитаны моделью заново, 560 полей сопоставлены с сохраненным результатом: совпали 535 (95,5 %), все 25 расхождений объяснились правилами объединения, ошибок чтения числа нет. Отдельно за аудит найдено и исправлено 10 дефектов кода объединения.

Разбор обозначений. Все 98 контрольных марок (40 из реальных проектов, остальные из бланков подбора и каталогов) разобраны верно.

5. Заключение

Схема решает поставленную задачу: модель надежно читает листы, детекторы находят листы ХОВС и спецификаций у новых проектировщиков

практически полностью, марки разбираются формальными грамматиками без модели. Схема работает в облачном сервисе подбора оборудования; по отзывам заводов, данные пригодны для подбора без ручного переноса.

Список литературы

[1] *ГОСТ 21.602-2016. СПДС. Правила выполнения рабочей документации отопления, вентиляции и кондиционирования*, Стандартинформ, М., 2016.

[2] B. T. Willard, R. Louf, "Efficient guided generation for large language models", 2023, arXiv: 2307.09702.

[3] В. Б. Кудрявцев, С. В. Алешин, А. С. Подколзин, *Введение в теорию автоматов*, Наука, М., 1985.

Статья поступила 8 ноября 2026 г.

Hybrid extraction of air handling unit configurations from building design documentation: formal grammars and large language models

V. V. Osokin, R. F. Alimov

HVAC equipment manufacturers transfer air handling unit data from building designs into selection software by hand. We describe a scheme for extracting this data automatically: a large language model reads the sheet image into a JSON schema and classifies sheets not recognised from text; the remaining sheet selection, parsing of unit designations by formal grammars, source merging and all decisions are done by verifiable code. Recall of system-table and specification sheets is 100 % on the training corpus and 94–99 % on a held-out sample; a recheck of 560 extracted fields found no number-reading errors.

Keywords: large language models, information extraction, formal grammars, design documentation, air handling units.

References

[1] *GOST 21.602-2016. SPDS. Rules for working documentation of heating, ventilation and air conditioning*, Standartinform, M., 2016 (in Russian).

[2] B. T. Willard, R. Louf, "Efficient guided generation for large language models", 2023, arXiv: 2307.09702.

[3] V. B. Kudryavtsev, S. V. Aleshin, A. S. Podkolzin, *Vvedenie v teoriyu avtomatov*, Nauka, M., 1985 (in Russian).

Received on November 8, 2026