**УДК 004.822**

# АВТОМАТИЗИРОВАННОЕ ИЗВЛЕЧЕНИЕ СЕМАНТИЧЕСКИХ СВЯЗЕЙ ИЗ КОРПУСА СТРУКТУРИРОВАННЫХ ТЕКСТОВ

**Ночевной Д.С.** (Университет ИТМО)
**Научный руководитель – старший преподаватель Клименков С.В.**
(Университет ИТМО)

Для построения современных онтологий всегда актуальным является извлечение семантически связанных слов из текста на естественном языке, однако общим недостатком таких онтологий является отсутствие специализированных терминов. Чаще всего онтологии дополняются специализированными терминами путем ручного занесения узлов и семантических связей, но этот подход является малоэффективным. В рамках научной работы составлена модель обработки документов в унифицированном формате, а также сформулирован подход для преобразования структурированных и слабоструктурированных текстов на естественном языке в информационные узлы семантической сети, отличающийся использованием грамматических конструкций и семантически значимых элементов форматирования, обеспечивающий структуризацию и декомпозицию данных.

Одним из ключевых элементов систем автоматической обработки текста являются онтологии или тезаурусы. Обычно в этом качестве используются онтологии или тезаурусы, построенные на базе тех или иных словарей. Однако общим недостатком таких онтологий является отсутствие специализированных терминов, специфичных для данной предметной области.

Чаще всего онтологии дополняются специализированными терминами путем ручного занесения узлов и семантических связей. Но этот подход по определению является малоэффективным и весьма затруднительным – создание онтологии достаточно большого объема и ее постоянная поддержка в актуальном состоянии требует огромных временных и финансовых затрат. Очевидно, что ручной сбор данных для онтологий – утомительная и громоздкая задача в виду очень большого объема этих данных и количества связей между ними. Более того, подобная деятельность требует глубокого понимания предметной области. Вследствие этого факта, во множестве случаев полученный результат может не иметь требуемой точности и полноты. Поэтому возникает проблема дополнения существующей онтологии узлами и связями из внешних источников.

Целью работы стало повышение точности и полноты информации в исходной онтологии, в частности, добавление новых связей типа «меронимия». Для достижения поставленной цели были выявлены следующие задачи, требующие решения:

1. поиск меронимов и холонимов в различных источниках неструктурированной и слабоструктурированной текстовой информации;
2. анализ найденных элементов;
3. добавление узлов и связей в исходную онтологию.

В рамках научного исследования рассмотрено множество существующих решений в данной области, а именно:

1. основанные на шаблонах (к примеру, PART “part of” WHOLE для связи часть-целое);
2. основанные на объектной модели документа (часто вложенные и родительские элементы DOM имеют семантическую связь);
3. основанные на форматировании текстовой информации (в частности, списки);
4. основанные на машинном обучении.

В наших предыдущих исследованиях были описаны и реализованы метод, основанный на анализе структуры текстовой информации, а также метод, основанный на машинном обучении с использованием инструмента «Word2vec». С помощью этих подходов были получены достаточно точные результаты при анализе форматирования текстовой информации. Однако они не учитывали грамматическую структуру текстовой информации, являющуюся не менее значимой, чем форматирование.

Исходя из всего вышесказанного, была предложена новая модель обработки текста с использованием документов в специализированном унифицированном формате, позволяющая единообразно работать с текстами, полученными из множества

структурированных и слабоструктурированных источников. Унифицированный формат предполагает хранение не только самого текста, но и различных атрибутов форматирования, а также грамматических конструкций.

Для выполнения поставленных задач был использован корпус текстов русского языка, имеющих специальную разметку с указанием частей речи, падежей и других характеристик каждого слова в предложении. Именно эта информация и стала ключевой для принятия решения о добавлении узлов и связей в исходную онтологию. В результате предложен подход для преобразования структурированных и слабоструктурированных текстов на естественном языке в информационные узлы семантической сети, отличающийся использованием грамматических конструкций и семантически значимых элементов форматирования, обеспечивающий структуризацию и декомпозицию данных.

Основные промежуточные результаты работы.

1. Рассмотрены основные подходы и методы, существующие в современном мире, для извлечения семантических связей из структурированных и слабоструктурированных источников информации.
2. Составлена модель обработки текстовых документов в специализированном унифицированном формате.
3. Сформулирован подход для преобразования структурированных и слабоструктурированных текстов на естественном языке в информационные узлы семантической сети, отличающийся использованием грамматических конструкций и семантически значимых элементов форматирования, обеспечивающий структуризацию и декомпозицию данных.