ВЫДЕЛЕНИЕ ПОНЯТИЙ ДЛЯ ФОРМИРОВАНИЯ ПРИКЛАДНОЙ ОНТОЛОГИИ НА БАЗЕ ТЕКСТОВ ПРОТОКОЛОВ ДОПРОСОВ

Н.В. Деева
Кафедра программного обеспечения интеллектуальных и компьютерных систем
Гродненский государственный университет имени Янки Купалы
Гродно, Республика Беларусь
E-mail: ndeeva@grsu.by

В данной статье рассматривается проблема извлечения смысловой информации из естественно-языковых текстов протоколов допросов досудебного производства. Анализируется класс уголовно-процессуальных документов, а именно протоколы допросов. Предлагается схема выделения понятий предметной области из материалов расследуемого происшествия, на базе которых предлагается строить прикладную онтологию дела.

Введение

Задача извлечения смысловой информации из естественно-языковых текстов сегодня действительно очень важная. В наш информационный век подавляющее большинство отраслей производственной деятельности человека компьютеризировано, огромные массивы бумажных носителей уступают место компактным электронным. Пока, к сожалению, нельзя сказать, что все оперативные работники, следователи, эксперты-криминалисты обеспечены персональными современными компьютерами. Однако завершенная в 2010 году программа «Электронная Беларусь» и открытая программа на 2011-2015 годы «Электронная Беларусь 2» однозначно убеждают нас в том, что изучение электронных документов с целью извлечения из них смысловой информации с последующим ее анализом и обобщением действительно актуальна.

На данный момент существует множество программных инструментов, находящихся в том числе в открытом доступе, для построения аннотаций, реферата, списка ключевых слов по тексту на естественном языке. Большинство интернет-поисковиков работают, используя алгоритмы и методы компьютерной лингвистики. Все еще набирает силу идея «семантического веба», призванная упорядочить и систематизировать информацию в Интернете, предложенная консорциумом W3C еще в конце 90-х годов XX века. Однако задача извлечения смысла из текстовых документов до сих пор не решена в полной мере в виду огромной сложности формализовать естественные языки, которые представляют собой живые организмы, развивающиеся во времени и даже пространстве.

В данной статье предлагается алгоритм выделения набора понятий для дальнейшего построения прикладной онтологии расследуемого происшествия по естественно-языковым описаниям в уголовно-процессуальных документах. Новизна данного подхода обуславливается невозможностью формирования онтологии экспертами или построением ограниченных прикладных онтологий до момента возникновения самого происшествия и проведения следственно-розыскных мероприятий. Построение такого описания предметной области расследуемого дела позволит выполнять поддержку принятия решений следователем или группой следователей, динамически формировать общую картину происшествия по собранным на данный момент доказательствам и фактам, а также делать предположения о несоответствиях в показаниях или версиях.

I. Анализ структуры протокола допроса

Анализ перечня документов из [1], которые исследуются, обобщаются и создаются в ходе следственного процесса и формируют уголовное дело, позволил выполнить их классификацию и выделить два класса: документы по ходу расследования и по сути расследования [2]. Согласно заявленным ранее целям исследования был выбран класс документов, содержащий наиболее актуальную для анализа следователем, динамически изменяющую общее представление о деле, информацию по сути расследуемого дела – протокол. Протокол можно определить, как акт, составленный уполномоченными на это должностными лицами (судебными или административными) для удостоверения тех или иных событий. Согласно проведенному исследованию протоколов были выделены следующие типы: заявления, явки, допросы, осмотры, обыски, задержания, извлечения, освидетельствования, следственные эксперименты, очные ставки, предъявления, проверки показаний.

В данной работе особое внимание уделяется анализу содержания протокола допроса. Согласно исследованиям в [2], протокол допроса состоит из следующих сегментов: статическая формулировка, повествовательный и вопросно-ответный. Первый сегмент содержит сведения о субъекте(ах) и объекте допроса, а также включа-

ет в текст допроса выдержки из Кодексов и статические формулировки. Наибольший интерес представляет изучение стилистических характеристик повествовательного и вопросно-ответного сегментов, так на основании исследований Татарниковой Н.М. [3], текст указанных сегментов характеризуется, как правило, линейным изложением фактов, упорядоченных по времени. Сегменты могут содержать статические фрагменты-описания, референты и нацелены на отражение реальных фактов. Такое описание соответствует официально-деловому стилю юрисдикциональных текстов, которые характеризуются как информативный нарратив-воспоминание с перфектной перспективой диктума. Согласно исследованиям [4], которые проводились над протоколами допросов, большую часть лексики такого типа документов составляют стилистически нейтральные слова: имена существительные, обозначающие конкретные предметы – до 40%; имена прилагательные, как правило, качественные – до 4%; стоп-слова (предлоги, союзы, частицы), в среднестатистическом диапазоне – 15-20%. Тексты характеризуются также перегрузкой числительными, в основном представленных в числовом формате.

II. Схема анализа текста протокола допроса

Исходя из определенных выше свойств повествовательного сегмента протокола допроса было принято решение проанализировать текст этого сегмента по следующей схеме. Выполним нормализацию текста, исключив из него лишние пробельные символы, а также выполним поиск с последующей заменой найденных данных на специальную конструкцию <дата>, <время>, <иное числовое значение>. Поиск временных определений выполняем с помощью теории исчисления предикатов и продукционных моделей. В результате можно исключить из дальнейшей обработки уже найденные определения. Пусть дан набор n нормализованных текстов $T = \{t_1, \ldots, t_n\}$ по одной предметной области (делу). Для каждого из них проведем лексический анализ и получим набор текстов с выделенными на них коллекциями абзацев p, предложений s, слов w и специальных конструкций и знаков пунктуации o: $T^l = \{t_1^l, \ldots, t_n^l\}$, где $t_i^l =< \{p\}, \{s\}, \{w\}, \{o\} >$.

Каждый t_i^l представляет собой иерархическую структуру $t_i^l =< V, E >$, где первому уровню дерева V^1 соответствует абзац, второму V^2 – предложение, а третьему V^3 – слово или специальная конструкция и знак пунктуации. Корнем дерева является весь текст, ребра дерева соответствуют отношению «часть-целое». Лексический анализ выполняем построением детерминированного конечного автомата с цепочкой символов текста протокола на входе. Затем для каждого терминального узла дерева, если оно является словом, проведем его морфологический анализ, то есть припишем ему наборы морфологических признаков (граммем), учитывая возможное появление омоформ. Получим следующее представление каждого текста в наборе: $t_i^l M =< \{p\}, \{s\}, \{w^M\}, \{o\} >$, где $w^M = \{m_1, \ldots, m_c\}, c \to 1$. Нахождение морфологических характеристик выполняем с помощью COM-объекта рабочей группы АОТ, размещенный в открытом доступе по адресу `http://aot.ru/`.

На следующем этапе для каждого текста выделим претендентов для формирования списка понятий прикладной онтологии по расследуемому факту. Будем считать, что в число претендентов попадают те слова, которые однозначно определены как существительные, либо если можно разрешить морфологическую омонимию нахождением в списке контактно-стоящих слов, согласованных по морфологическим характеристикам. Таким образом, получим наборы вероятных понятий для каждого текста $D' = \{d_1, \ldots, d_n\}$, где $d_i =< a_i^1, \ldots, a_i^{k_i} >$. А искомый набор понятий получаем пересечением этих множеств: $D = \cap d_i, 0 < i \leq n$.

Таким образом, построено множество понятий, которое может быть использовано в качестве базы построения прикладной онтологии по одному происшествию.

Заключение

Полученные результаты могут быть использованы для поверхностного анализа массивов протоколов допросов, для последующих исследований в области построения прикладных онтологий, а также поиска несоответствий в показаниях фигурантов расследуемого дела.

1. Сборник образцов уголовно-процессуальных документов с комментариями. Возбуждение уголовного дела и предварительное расследование: учебн. – практ. пособие / авт.–сост. Г. Н. Васильев, М. И. Емшанский, Е. И. Климова [и др.]; под рук. и науч. ред. проф. М. А. Шостака. – Мн.: Амалфея, 2006. – 704 с.
2. Буза, М. К. Информационная модель процесса расследования правонарушений / М. К. Буза, Н. В. Деева // Информатика – 2009. – №4(24). – С. 112–123.
3. Татарникова, Н. М. Стилевая черта vs коммуникативная стратегия (на материале жанров юрисдикционного подстиля) / Н. М. Татарникова // Проблемы концептуализации действительности и моделирования языковой картины мира: сб. науч. тр.: вып. 4 / сост., отв. ред. Т. В. Симашко. – М.; Архангельск, 2009. – С. 90–95.
4. Кыркунова, Л. Г. Функционально-смысловые типы речи в аспекте внутристилевой дифференциации следственно-судебных текстов / Л. Г. Кыркунова // Стереотипность и творчество в тексте – 2004. – Вып.7. – С. 290–312.