OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

АВТОМАТИЗАЦИЯ СЕМАНТИЧЕСКОГО АНАЛИЗА НОВОСТНЫХ ИНТЕРНЕТ-ТЕКСТОВ

Солошенко А.Н., Розалиев В.Л., Заболеева-Зотова А.В.

Волгоградский государственный технический университет,
г. Волгоград, Россия

nastyasolan@gmail.com

vladimir.rozaliev@gmail.com

zabzot@gmail.com

Данная статья посвящена проблеме представления неструктурированных новостных сюжетов в сжатом виде с сохранением их смысла. Особое внимание уделено интеграции систем с новостными сайтами и социальными сетями. В работе скомбинированы статистические алгоритмы извлечения ключевых слов и алгоритмы формирования семантической связности блоков текста.

Ключевые слова: семантика, синтаксический анализ, новостной интернет-текст, интеграция с новостными сервисами.

Введение

Развитие информационных ресурсов Internet многократно усилило проблему информационной перегрузки. Еще в начале XXI века американская исследовательская служба Cyveillance сообщила о том, что количество страниц в Internet превысило 4 млрд, и с каждым днем увеличивается на 7 млн. «Сырые» неструктурированные данные составляют большую часть информации, с которой имеют дело пользователи, поэтому многие организации (службы рассылки новостей, информационно-библиотечные системы и др.) и частные лица заинтересованы в эффективных технологиях автоматизированного семантического анализа информации, представленной на естественном языке [Ландэ Д.В., 2005; Розалиев и др., 2010] .

При этом необходимо отметить, что темпы роста аудитории онлайновых новостных ресурсов практически вдвое превышают темпы роста общей численности пользователей интернета, и к сегодняшнему дню данная аудитория составляет 43,2% российских интернет-пользователей (согласно исследованию, проведенному сотрудниками Nielsen//NetRatings). Так, недельная аудитория составляет в среднем 6,3 миллиона пользователей: новостных сайтов – 3,9 млн., газет – 3,7 млн., информационных агентств – 2,2 млн., радиостанций и телеканалов – 1,3 миллиона. Число ежедневных сообщений в Twitter также приблизилось к отметке 400 млн. записей в день, тогда как в апреле 2012 года этот показатель составлял 340 млн. в сутки.

Данные факты говорят о том, что необходимость использования разрабатываемой системы, позволяющей проанализировать и представить информацию в сжатом виде, но с сохранением смысла, с каждым годом будет возрастать. Существующие программные системы полностью не решают данную проблему. Причиной этому является отсутствие моделей и методов, обеспечивающих формализацию и адекватный семантический анализ текстов на естественном языке. Объяснением этого является сложность и неоднозначность решения задачи семантического анализа для различного вида текстов. Кроме того, большинство систем не направлены на обработку новостных текстов.

Основная идея заключается в разработке нового подхода к обработке текстовой информации, основанного на комбинировании нескольких типов подходов: традиционного статистического и более сложного лингвистического.

1. Обзор существующих систем семантического анализа новостных текстов

На международном рынке представлено множество программных продуктов, которые позволяют проанализировать текст с точки зрения семантики.

Среди отечественных стоит выделить АОТ и Semantic Analyzer Group, из зарубежных – мощный инструмент анализа текстов IBM Text Miner. Технология Semantic Analyzer Group, как и АОТ (Автоматическая Обработка Текста) позволяет строить синтактико-семантическую сеть, включает графематический, морфологический и семантический модули. Обе системы работают со словарями и тезаурусами. IBM Text Miner содержит утилиты классификации, кластеризации, поиска ключевых слов и составления аннотации текстов. Однако на обработку новостных статей программы не направлены. [Заболеева-Зотова и др., 2010a]

Российская система Яндекс Новости позволяет автоматически группировать данные в новостные сюжеты и составлять аннотации статей на основе кластера новостных документов. Сервис InfoStream, обеспечивает доступ к оперативной информации в поисковом режиме с учетом семантической близости документов.

Прямым аналогом системы является мобильный агрегатор новостей Summly, купленный в марте 2013 компанией года Yahoo!. Summly – приложение под iPhone, которое позволяет сжимать произвольную статью в резюме до 400 знаков и подбирать подходящие картинки для оформления на экране мобильного устройства. Однако приложение абсолютно неприменимо для обработки текстов на русском языке.

Таким образом, разрабатываемая система призвана устранить недостатки существующих систем, модифицировать существующие решения семантического анализа применительно к новостным текстам на русском языке.

2. Интеграция с новостными сервисами

Для удобства пользователей в разрабатываемой системе заложена возможность получения необходимой статьи с новостного сайта по ее URL, т.е. будет реализована интеграция с наиболее популярными новостными сайтами (например, rbc.ru, newsru.com, expert.ru, kommersant.ru, lenta.ru). Для этого разрабатывается лексический анализатор новостных сайтов, при помощи которого из разметки HTML можно получить необходимый текст.

Далее рассмотрим решение задачи семантического анализа новостного текста по этапам.

3. Методика семантического анализа новостного текста

Решение задачи семантического анализа можно разбить на несколько этапов: предварительная обработка текста (графематический и морфологический анализ), синтаксический анализ и затем непосредственно семантическая обработка.

В основе структуры новости положен принцип «перевернутой пирамиды»: заголовок отражает тему и содержит не более 10 слов, основные факты, касающиеся события, отражены в 1-2 абзацах (лид), 3-й и последующие абзацы составляют бэкграунд (контекст) [Солошенко А.Н. и др., 2013]. В задачу графематического анализа входит внутреннее представление структуры новости: T = <P, S, W>, где P – абзацы, S – предложения, W – слова. При этом необходимо корректно выделить заголовок и первое предложение абзаца, содержащее основные факты статьи. Для морфоанализа целесообразно использовать словарные методы, за основу можно взять, например, словарь А.А. Зализняка.

В простейшем случае структуру совокупности знаний S текста новости можно определить следующим образом: S={M,F}, где M – множество всех понятий данной совокупности знаний, F – отношение «смысловая связь». В качестве формальной модели структуры знаний можно использовать семантическую сеть, определяемую в виде ориентированного графа G = (E,V), где E - множество вершин, поставленное во взаимно однозначное соответствие с множеством понятий; V - множество ориентированных дуг; дуга выходит из вершины, соответствующей основному понятию A, и входит в вершину, соответствующую понятию, которое сочетается по смыслу в тексте с понятием A [Заболеева-Зотова А.В. и др., 2010b; Михайлов Д.В. и др., 2009].

Для построения вышеописанного семантического графа, или семантической сети, необходимо проведение синтаксического анализа, задача которого - выделение синтаксических конструкций, определение связности и подчинения фрагментов. Для поиска фрагментов потребуется шаблон поиска - кортеж <N,S,P>, где N – нормальная форма искомого слова, S – часть речи и P={p} – множество искомых параметров искомого слова.

После нахождения фрагмента необходимо его преобразование.

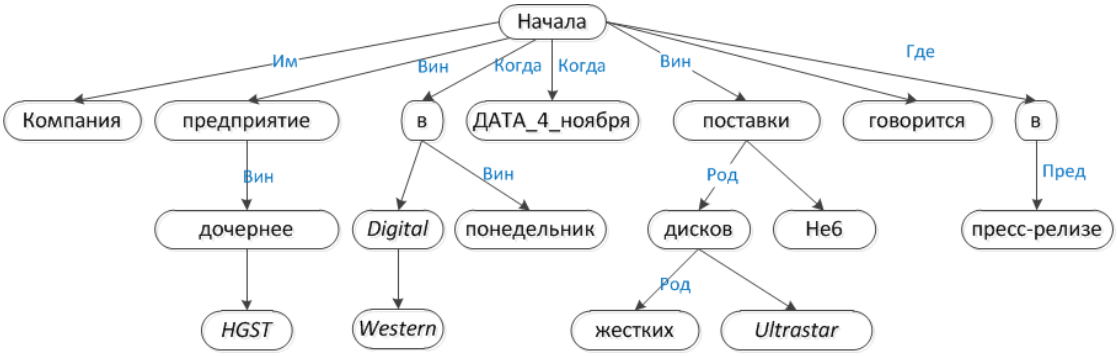

Рисунок 1 – Дерево зависисмостей предложения (семантический граф)

Таким образом, правило синтаксической сегментации будет состоять из шаблона поиска фрагмента, шаблона формирования фрагмента и списка исключений. В результате синтаксического анализа получается граф, подобный представленному на рисунке 1 [Большакова Е.И. и др., 2011].

После построения графа, можем выделить наиболее часто встречающиеся в тексте новости субъекты и объекты повествования, построить наглядное облако новости (рисунок 2, размер шрифта отражает важность ключевой фразы), аннотацию.

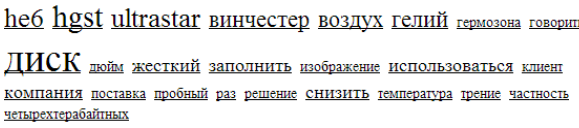

Рисунок 2 – Облако текста

При размещении статей в новостной ленте требуется определять рубрику, к которой относится анализируемая статья. Для этого целесообразно использовать метод латентно-семантического анализ [Машечкин И.В. и др., 2011; Тарасов С.Д., 2008]. На первом шаге требуется составить частотную матрицу индексируемых слов.

Следующим шагом мы проводим сингулярное разложение полученной матрицы. Сингулярное разложение – это математическая операция раскладывающая матрицу на три составляющих. Т.е. исходную матрицу M мы представляем в виде:

$$M = U \cdot W \cdot V^t \qquad (1)$$

Где U и V^t – ортогональные матрицы, а W – диагональная матрица. Причем диагональные элементы матрицы W упорядочены в порядке убывания. Диагональные элементы матрицы W называются сингулярными числами.

Достоинство сингулярного разложения состоит в том, что оно выделяет ключевые составляющие матрицы, позволяя игнорировать шумы. Согласно простым правилам произведения матриц, столбцы и строки, соответствующие меньшим сингулярным значениям, дают наименьший вклад в итоговое произведение. Важно, что при этом гарантируется, оптимальность полученного произведения. Таким образом гарантируется достаточное точное определение тематики новостной статьи.

4. Пример применения системы

Рассмотрим пример работы системы на примере новостного сообщения, взятого из Интернет, заглавием которого является строка «Пользователи из США предпочли карты Apple картам Google» (рисунок 4).

Предполагается возможность настройки агрегации новостей по темам или ключевым словам. Для подобранных текстов, кроме аннотирования и построения деревьев зависимости (описанных выше), доступна функция комплексного анализа текста (определение темы, ключевых сущностей, синтаксических и морфологических характеристик), цитирования новостей, сокращенных до определенного формата, в социальных сетях (например, Твиттере). Возможно сохранить понравившиеся обзоры.

Был проведен ряд экспериментов, в ходе которых определено, что время, затраченное на обработку текста вручную составляет 2,5 минут, время обработки с помощью программы – порядка 10 секунд (рисунок 3).

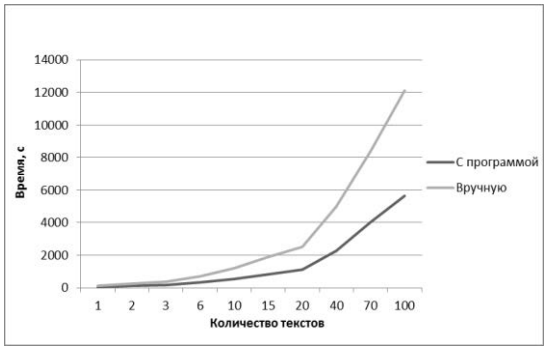

Рисунок 3 – Зависимость времени обработки от количества текстов

Рисунок 4 – Пример работы системы

Заключение

Таким образом, совмещение статистических и лингвистических алгоритмов семантического анализа позволит улучшить качество обработки новостных статей для последующей публикации в новостных лентах. В условиях нынешнего информационного века, с огромным количеством новостных сообщений в сети интернет, применение таких технологий является необходимостью, так как значительно ускоряет обработку повседневной информации.

Работа частично поддержана Российским фондом фундаментальных исследований (проекты 12-07-00266, 12-07-00270, 13-07-00459, 13-07-97042).

Библиографический список

[Большакова Е.И. и др., 2011] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие. / Большакова Е.И. [и др.]. – М. : МИЭМ, 2011.

[Заболеева-Зотова и др., 2010a] Автоматизация семантического анализа текста технического задания: монография. / Заболеева-Зотова А.В., Орлова Ю.А. - Волгоград, ИУНЛ. 2010. - 155 с.

[Заболеева-Зотова и др., 2010b] Автоматизация начальных этапов проектирования программного обеспечения / Заболеева-Зотова А.В., Орлова Ю.А. // Изв. ВолгГТУ. Серия «Актуальные проблемы управления, вычислительной техники и информатики в технических системах»: межвуз. сб. науч. ст. - Волгоград, ВолгГТУ. 2010. - Вып. 8, № 6. - С. 121-124.

[Ландэ Д.В., 2005] Ландэ, Д. В. Поиск знаний в INTERNET. Профессиональная работа. : Пер. с англ. – М. : Диалектика, 2005. – 272 с. : ил.

[Машечкин И.В. и др., 2011] Машечкин, И.В. Латентно-семантический анализ в задаче автоматического аннотирования / Машечкин И.В., Петровский М.И. // Программирование. – 2011. – Т. 37, № 6. – 67-77.

[Михайлов Д.В. и др., 2009] Михайлов, Д. В. Морфология и синтаксис в задаче семантической кластеризации / Михайлов Д. В., Емельянов Г. М. // Математические методы распознавания образов (ММРО-14), Владимирская область, Суздаль, 21-26 сентября 2009 г. – 2009.

[Розалиев и др., 2010] Моделирование эмоционального состояния человека на основе гибридных методов / Розалиев В.Л., Заболеева-Зотова А.В. // Программные продукты и системы: международный науч.-практ. журнал. – Тверь, 2010 – Вып.2 (90). – С.141-146.

[Солошенко А.Н. и др., 2013] Солошенко, А.Н. Автоматизированное составление обзорных рефератов новостных Интернет-текстов / Солошенко А.Н., Орлова Ю.А., Розалиев В.Л., Заболеева-Зотова А.В. // Труды Конгресса по интеллектуальным системам и информационным технологиям "IS&IT`13", п. Дивноморское, 3-9 сент. 2013 г. В 4 т. Т. 1 / Рос. ассоциация искусственного интеллекта, ФГАОУ ВПО «Южный федеральный ун-т» [и др.]. - М., 2013. - С. 233-238.

[Тарасов С.Д., 2008] Тарасов, С. Д. Алгоритм ранжирования связных структур в задачах автоматического составления обзорных рефератов новостных сюжетов.// RuSSIR'2008, труды Второй Российской конференции молодых ученых по информационному поиску. – Таганрог: Изд-во ТТИ ЮФУ, 2008. – С. 90-100.

AUTOMATION OF SEMANTIC ANALYSIS OF INTERNET NEWS TEXTS

Soloshenko A.N., Rozaliev V.L., Zaboleeva-Zotova A.V.

Volgograd State Technical University, Volgograd, Russia

nastyasolan@gmail.com

vladimir.rozaliev@gmail.com

zabzot@gmail.com

Semantic analysis of the text received in recent years a considerable urgency in connection with development of Internet and catalogs of information resources. This article is devoted to a problem of unstructured news stories representation in compressed form. Particular attention is paid to the integration of systems with news services and social networks, text parsing, developing of models and methods of news texts semantic analysis based on a combination of statistical algorithms for extracting keywords and algorithms forming the semantic coherence of text parts.

Keywords: semantics, syntactic analysis, Internet news texts, integration with news services.