ТЕМАТИЧЕСКИЙ АНАЛИЗ ЕСТЕСТВЕННО ЯЗЫКОВЫХ ТЕКСТОВ

THEMATIC ANALYSIS OF NATURAL LANGUAGE TEXTS

В.П. Гладун *(glad@aduis.kiev.ua),*

В.Ю. Величко,

Л.А. Святогор

Институт кибернетики им. В.М. Глушкова НАН Украины

Статья посвящена проблеме создания сжатого образа текста. Некоторые применения естественно-языковых текстов требуют такой формы представления текста, которая была бы результатом разумного компромисса между желанием сделать текст короче с сохранением основных тематических целей и желанием сделать исходный текст наиболее полным. Каковы пути этого компромисса? В докладе обсуждается эта проблема. Приведенный метод реализован в программной системе KONSPEKT.

Введение

В поисковых системах, обрабатывающих естественно-языковые тексты, тематика текстов обычно определяется ключевым словом или словосочетанием. Возникают трудности, вызванные тем обстоятельством, что тексты чаще всего не являются однотемными, но представляют собой переплетение нескольких связанных или независимых тем. Использование в этих случаях в целях тематического анализа только одного ключа оказывается недостаточным. Необходимо создать механизм, который бы улавливал тематические повороты и позволял отслеживать изгибы темы на всех этапах анализа текста. В статье в качестве такого механизма предлагается онтология ассоциаций.

Указанный подход реализован в программной системе KONSPEKT (Институт кибернетики им. В.М. Глушкова НАН Украины). Система формирует сокращённый пересказ исходного естественно-языкового текста путём отбора из исходного текста предложений, удовлетворяющих определённым критериям. Полученный пересказ удобно назвать конспектом (англ. synopsis).

Конспект является формой представления текстов, в которой объединяются два противоречивых требования к обобщенному представлению текстовых знаний: 1) сокращение объема текста при сохранении основных тематических целей (реферирование, аннотирование); 2) пересказ основных положений содержания при сохранении связности текста.

По отношению к исходному тексту конспект должен быть результатом разумного компромисса между желанием сократить его объем, жертвуя некоторыми элементами содержания (например, в целях экономии памяти или совершенствования поисковых операций), и желанием как можно полнее передать детали содержания. При удачном разрешении этого компромисса конспект становится наиболее рациональной формой представления текстов. Каковы пути выполнения указанных требований? Совершенно очевидной является необходимость использования средств синтаксического и семантического анализа.

Семантический анализ текста

Все фразы исходного текста подвергаются синтактико-семантическому анализу. Основной операцией синтактико-семантического анализа является распознавание синтаксических и семантических отношений, связывающих слова текста.

Распознавание синтаксических и семантических связей между знаменательными словами осуществляется путем анализа флексий и предлогов без использования категорий и правил традиционной грамматики. По результатам синтактико-семантического анализа в исходном тексте отбираются предложения, содержащие "ядерные конструкции".

Термин “ядерные конструкции” используется в трансформационной грамматике для обозначения простого базового суждения, путем трансформации которого формируется предложение в целом. В данном случае ядерной конструкцией служит предложение, состоящее из подлежащего, сказуемого и соединяющей их связки. Алгоритм синтактико-семантического анализа на основе лексических моделей описан в [1-4].

Таким образом, на этом этапе из исходного текста для дальнейшего анализа отбираются полносоставные предложения. Для каждого из отобранных предложений формируется n - шаговое расширение ядра – часть предложения, содержащая его ядро, а также слова, связанные в дереве зависимостей с элементами ядра путями, длина которых не превышает n (заданный параметр).

В результате синтактико-семантического анализа формируется массив n -шаговых расширений полносоставных предложений исходного текста. Далее осуществляется тематический анализ текста. Тематический анализ текста выполняется на основе онтологии ассоциаций.

Онтология ассоциаций

В дальнейшем будем пользоваться следующими определениями. Ассоциация понятий представляет собой множество понятий, имеющих общую семантическую характеристику (ассоциативный признак). Ассоциативным признаком может быть вид деятельности или отдельное действие, с которым связано понятие, причастность понятия к какому-либо типу событий, явлений, ситуаций, (например, свадьба, рождество, выборы и т.п.), временной период или интервал (зима, лето, июль и т.п.).

В данной разработке используются три группы ассоциаций: A (action)-по виду деятельности; S (situation)-по типу ситуаций; T (time)-по времени.

В каждой группе ассоциаций выделяются центры ассоциаций – понятия, вокруг которых группируются ассоциации. Термины, обозначающие виды деятельности или действия, входят в ассоциации группы ***A***, с центрами которых их связывает отношение «подвид». Например, термины «искусственный интеллект», «принятие решений», «извлечение знаний» входят в ассоциацию «информатика», так как они обозначают разделы этого вида деятельности.

Онтология ассоциаций – словарь терминов, в котором термины индексированы сокращёнными обозначениями центров ассоциаций, в которые они входят.

При индексах ассоциаций записывается целое число, указывающее "вес" ассоциации.

Термины, обозначающие объекты, входят в ассоциации группы A, если они связаны с каким-либо видом деятельности или действием отношением типа "семантический падеж" ("агент", "объект", "инструмент", "среда" и т.п.) Соответствующие индексы терминов обычно имеют вес 2. Например, термины "прораб", "топор", "данные" получают индексы соответственно ст2, ст2, ин2. Это означает, что они входят в ассоциации "строительство" или "информатика" с весом 2.

Связь термина с ассоциациями групп S и t указывается путем ввода в индекс термина сокращенных обозначений ассоциаций, входящих в эти группы. Обычно таким ассоциациям присваивают веса 2 или 3. Например, в индекс термина "елка" могут входить обозначения ро3 (рождество), зи2 (зима).

Индексирование терминов в онтологии ассоциаций выполняется экспертами и, естественно, отображает их субъективное понимание индексируемых терминов. Для удобства выполнения индексирования в распоряжении экспертов должна быть некоторая классификация видов деятельности и отдельных действий, а также список характерных классов событий, явлений, ситуаций и привычных терминов, именующих временные периоды. Эта информация используется для выбора центров ассоциаций.

Тематический анализ текста

Для поддержки связности конспекта авторы предлагают подход, при котором текст рассматривается как совокупность текстовых фрагментов, раскрывающих отдельные связанные или независимые темы. Процесс тематического анализа организован циклически. На каждом прогоне цикла из исходного текста в формируемый конспект отбираются предложения, **n**-шаговые расширения которых содержат ключевой термин, который при запуске цикла задаётся пользователем системы, а на последующих его прогонах формируется автоматически. В качестве очередного ключа в оставшихся (еще не отобранных в конспект) предложениях выбирается термин, имеющий наибольший суммарный вес ассоциаций, которые являются общими для этого слова и ключа, используемого на предыдущем прогоне цикла. Таким образом осуществляется упорядоченный перебор тематики исходного текста, причем за счет выбора в качестве очередного ключа знаменательного слова, наиболее близкого по ассоциациям предыдущему ключу, каждая очередная тема оказывается связанной с предыдущей. В результате формируется сокращенный текст, который по объему и связности близок к сложившимся представлениям о свойствах конспектов.

Механизм выбора нового ключа на основе онтологии ассоциаций в какой-то степени является моделью ассоциативного мышления, что делает данную разработку, по мнению авторов более перспективной.

Система KONSPEKT

Система КОНСПЕКТ выполняет выделение и сжатое конспектирование исходных естественно-языковых текстов, относящихся к заданной теме, которая задается ключевым словом или словосочетанием.

Система предоставляет возможность при выборе текстов обращаться в Интернет или задавать исходные тексты в локальных файлах. При обращении в Интернет нужно выбрать поисковую машину. Пользователь задаёт первое ключевое слово, с которого начинается тематический анализ текста, и число ключей, которые может сгенерировать и использовать система в процессе тематического анализа.

На рис.1 в качестве примера приведен конспект данной статьи, построенный системой. В качестве исходного ключевого слова было указано слово «конспект». При анализе текста статьи система выбрала из онтологии ассоциаций последовательность, состоящую из 7 (заданный параметр) ключей и создала конспект, отобрав из исходного текста 27 предложений, значительно таким образом сократив объём полученного пересказа (27 предложений) по сравнению с объёмом исходного текста (103 предложения). Более подробный тематический

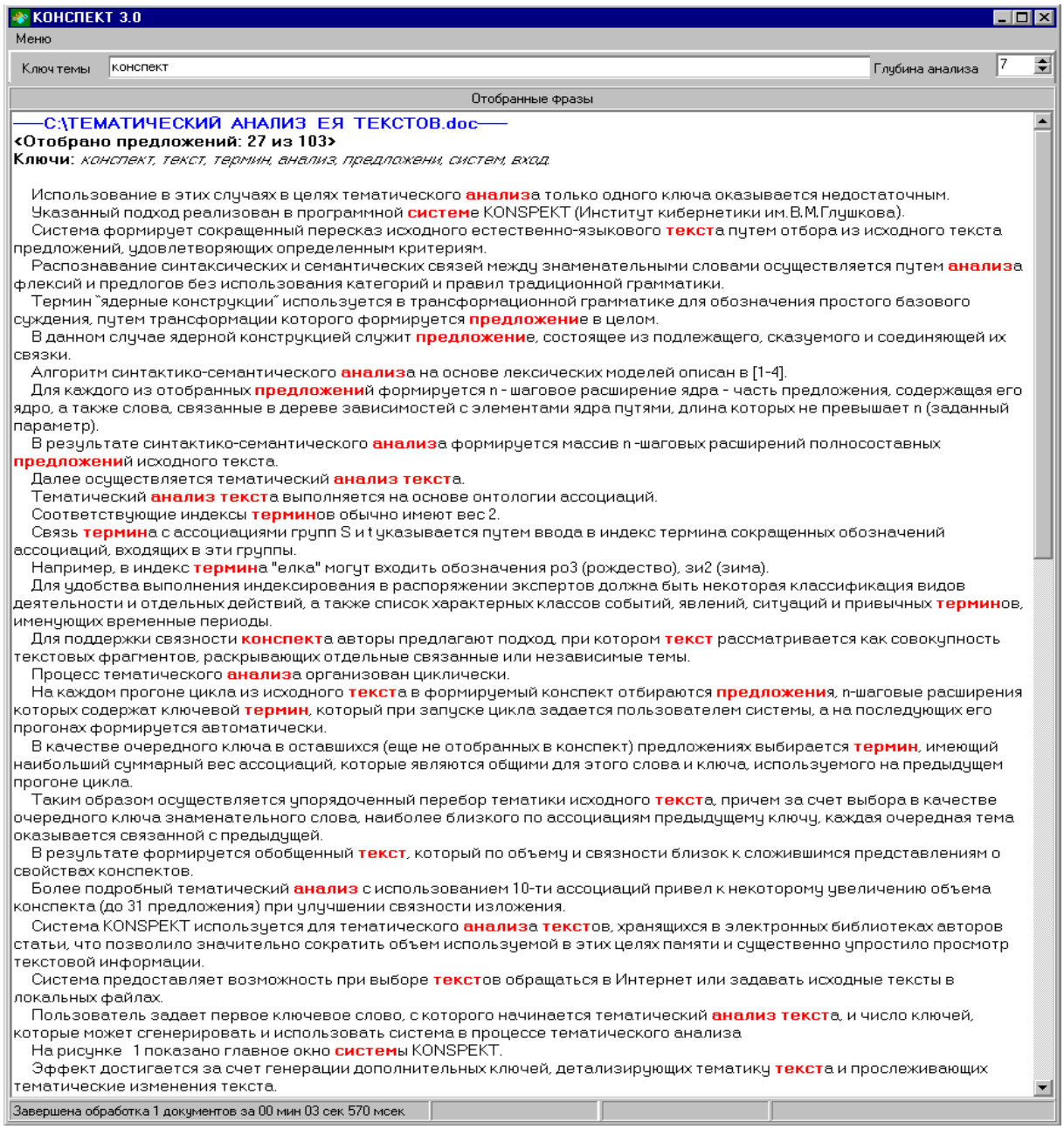

Рис.1 Пример работы системы KONSPEKT при глубине анализа 7.

анализ с использованием 10-ти ассоциаций привёл к некоторому увеличению объёма конспекта (до 31 предложения) при улучшении связности изложения.

Система KONSPEKT используется для тематического анализа текстов, хранящихся в электронных библиотеках авторов статьи, что позволило значительно сократить объём используемой в этих целях памяти и существенно облегчить просмотр текстовой информации.

Онтология ассоциаций позволяет эффективно реализовать различные процессы тематического анализа и синтеза естественно-языковых текстов.

В настоящее время онтология ассоциаций включает 106 ассоциаций. Проведенные эксперименты показали, что уже этот набор ассоциаций даёт возможность строить с помощью системы KONSPEKT краткое связное представление текста.

Разработка тематики продолжается по двум направлениям: 1) накопление опыта реального конспектирования и 2) обогащение онтологии ассоциаций. В следующем эксперименте был составлен автоматический конспект книги В.П. Гладуна «Партнёрство с компьютером» (Киев, 2000 г.). Исходный текст напечатан на 116-ти страницах и содержит 1477 предложений. Полный список онтологии ассоциаций включал 106 ассоциативных терминов. При настройке системы на поиск восьми ассоциаций в конспект было отобрано 248 фраз. После увеличения глубины поиска до пятнадцати ключей конспект увеличился до 384-х фраз.

Зависимость размера конспекта от числа используемых ассоциаций (ключей) представляется важной закономерностью, с помощью которой можно регулировать объём сжатого теста.

Заключение

Представленный в статье метод автоматического конспектирования естественно-языковых текстов позволяет формировать сжатые образы исходных текстов при сохранении основных положений авторского замысла. Эффект достигается за счёт генерации дополнительных ключей, детализирующих тематику текста и прослеживающих тематические изменения текста. С этой целью используется онтология ассоциаций. Механизм выбора нового ключа на основе онтологии ассоциаций в какой-то степени моделирует процесс ассоциативного мышления. Система KONSPEKT является эффективным инструментом при создании различных хранилищ текстовой информации.

Литература

1. Гладун В.П. Процессы формирования новых знаний // София: СД "Педагогь". 1994г. - 192с.

2. Гладун В.П. Планирование решений // Киев: Наукова думка, 1987. -168 с.

3. Гладун В.П. и др. Формирование тематических знаний на основе анализа ЕЯ текстов сети Интернет // Труды международной конференции Диалог'2003. М.: Наука, 2003. с. 190-192.

4. V.Gladun, A.Tkachev, V.Velichko, N.Vashchenko. Selection of Thematic NL-Knowledge from the INTERNET // International Journal "Information Theories & Applications" Vol.10 /2003 N.2 FOI-COMMERCE - Sofia.- pp.123-125.