

Исследование тематической структуры текста на основе большого лингвистического ресурса

Лукашевич Н.В., Добров Б.В.

Институт США и Канады; НИВЦ МГУ;

louk@mail.cir.ru, dobroff@mail.cir.ru

Введение

Ван Дейк и Кинч (Ван Дейк, Кинч, 1988) описывают тематическую структуру текста как иерархическую структуру в том смысле, что тема всего текста может быть обычно описана посредством более конкретных тем текста, которые в свою очередь могут быть охарактеризованы посредством еще более конкретных подтем и т.д. Каждое предложение связного текста посвящено раскрытию той или иной подтемы основной темы текста.

В настоящее время исследование тематической структуры текста приобрело большое практическое значение, поскольку нахождение механизмов автоматического определения основной темы и подтем документа могло бы значительно улучшить эффективность информационного поиска в информационных системах.

Формулировка основной темы текста содержит некоторую совокупность слов, наиболее значимых для передачи содержания текста. Если рассмотреть текст, то можно видеть, что слова и словосочетания, близкие по смыслу к словам основной темы, образуют лексические цепочки, которые пронизывают весь текст. Естественно предположить, что если имеется лингвистический ресурс, в котором описаны разнообразные смысловые связи между словами, то можно двигаться по тексту, находить связанные по смыслу слова, формировать лексические цепочки (Morris J., Hirst G. 1991). Самые частотные (или выделенные по другим критериям) цепочки могли бы показать, чему именно посвящен конкретный текст (Barzilay R., Elhadad M. 1997).

При формировании таких цепочек важнейшим является ответ на вопрос, какого рода смысловые отношения могут использоваться для организации связного текста. Наиболее известными способами лексической связности в тексте являются повторы слов, синонимы, видовые и родовые слова, а также другие виды связей, которые в разных источниках называются ситуативными (Halliday, Hasan 1976), “связями по фрейму” (Селезнев, 1987), консеквентами (Крылов, 1984) и т.п.

Данная статья посвящена рассмотрению различных смысловых связей между словами и терминами текста, которые используются в качестве инструмента тематической организации текста. Описание производится на основе большого семантико-

энциклопедического ресурса - Тезауруса русского языка, который позволяет описать такие смысловые связи как качественно, так и количественно, в терминах путей между понятиями в концептуальной сети. Важной особенностью использования заранее созданного ресурса для такого анализа текста является то, что он позволяет определить, насколько может варьироваться состав лексических цепочек в разных текстах. Например, на основе Тезауруса можно проверить, следует ли из сходства совокупностей слов, использованных в различных текстах, сходство их лексических цепочек.

2. Тезаурус русского языка как инструмент исследования связности текста

Исследования связного текста проводились нами в ходе развития и практического применения Тезауруса русского языка (35 тысяч концептов, 75 тысяч текстовых входов, более 130 тысяч прямых концептуальных отношений - (Лукашевич, 1999)), подчасть которого, называемая Общеполитическим тезаурусом (21 тысяча концептов, 38 тысяч текстовых входов более, 80 тысяч отношений между концептами - (Лукашевич, Салий 1997)) в течение ряда лет служила лингвистическим ресурсом для разнообразных практических приложений автоматической обработки текстов таких как автоматическое концептуальное индексирование, автоматическое рубрицирование и аннотирование больших потоков текстов. Терминология, описанная в Тезаурусе, с высокой степенью полноты покрывает термины официальных документов и прессы после 1991 года.

С точки зрения исследования механизмов лексической связности текста особенно важны следующие основные принципы описания концептуальных отношений в Тезаурусе:

- 1) Концептуальные связи Тезауруса с помощью специальных процедур постоянно проверяются в текущих практических приложениях с точки зрения того, насколько они отражают реальные лексические связи в том или ином тексте.
- 2) Набор основных концептуальных отношений в Тезаурусе весьма невелик. Это отношения ВЫШЕ-НИЖЕ, ЧАСТЬ-ЦЕЛОЕ, АССОЦИАЦИЯ. Однако именно этот небольшой набор связей позволяет сосредоточиться на установлении самой связи и избежать траты времени на поиск названия связи. Кроме того, по мнению авторов, использование большего содержательного разнообразия названий связей между понятиями при настоящем уровне понимания структуры текста не могут привести к каким-либо новым результатам при автоматической обработке больших массивов разнообразных текстов.
- 3) Для концептуальных отношений определены свойства наследования и транзитивности. Каждое концептуальное отношение не просто соединяет пару концептов, но и определяет направление движения по путям Тезауруса, что позволяет считать связанными по смыслу не только понятия Тезауруса, между которыми непосредственно установлена связь, но и некоторую более широкую совокупность понятий, соединенных с исходным понятием путями из нескольких отношений.
- 4) Большое внимание также уделялось описанию ситуативных связей понятий, для которых использовались связи ЦЕЛОЕ (*спортсмен* - ЦЕЛОЕ - *спорт*), симметричная ассоциация (*тракторист* - АСЦ - *трактор*), несимметричная ассоциация (*автомобиль* - АСЦ₂ - *парковка*; *парковка* - АСЦ₁ - *автомобиль*).

Полученный ресурс обладает значительными возможностями для восстановления картины смысловых связей, скрывающихся за текстом, анализа их структуры и способа использования для организации связного текста.

3. Формулировка темы, тематические узлы и лексические цепочки

Предположим, что мы сформулировали основную тему некоторого текста. В ней упомянуты некоторые понятия и/или конкретные объекты текста. Подтемы текста раскрывают взаимоотношения между этими основными понятиями/объектами и поэтому должны тем или иным образом ссылаться на них, используя повторы слов, синонимы или другие слова, семантически связанные с понятиями основной темы (далее *основные понятия* текста). Таким образом, основным понятиям текста соответствуют некоторые совокупности слов текста (и совокупность понятий, стоящими за этими словами), которые используются в данном тексте для ссылки на эти основные понятия.

Основное понятие текста и совокупность понятий, которые используются в данном тексте на ссылки на эти основные понятия в подтемах текста, назовем *основными тематическими узлами*. Соответственно, совокупности понятий, которые используются в качестве ссылок на понятия подтем документа, назовем *локальными тематическими узлами*.

Рассмотрим пример текста ("Финансовые известия", 1997, Константин Баскаев):

ГАЗПРОМ СТАНОВИТСЯ ФИНАНСОВОЙ ОСНОВОЙ ГОСУДАРСТВА

ТОЛЬКО благодаря Газпрому федеральный бюджет в июне выполнен на 100%. Концерн обеспечил 41% налоговых доходов госбюджета. Это текущие платежи Газпрома (2 трлн. руб.) и выплата задолженности (5,7 трлн. руб.). Всего в мае-июне компания выплатила в бюджет и пенсионный фонд 14,5 трлн. руб. Из них 11,5 трлн. руб. были сданы прямо в казначейство, "чтобы никакой банк не мог ими воспользоваться".

С такими результатами Газпром подошел к своему ежегодному собранию акционеров, которое состоялось 28 июня.

В 1996 г. чистая прибыль, оставшаяся в распоряжении Газпрома составила 33,2 млрд. руб. Капитализация общества на 1 июня составила 16,5 млрд. долл. То есть возросла в два раза за пять месяцев 1997 года. На каждую акцию по результатам 1996 года выплачивается 200% годовых. Годовое собрание акционеров 28 июня сочло эти результаты удовлетворительными.

Остается неясным вопрос с погашением задолженности потребителей газа, достигшей астрономической суммы - более 70 трлн. руб. Минобороны, части МВД, Российская академия наук и другие бюджетные организации должны Газпрому свыше 4 трлн. руб....

Для Газпрома на первом месте стоят интересы основного акционера - государства. Правлению общества и руководителям государства удалось прийти к общей точке зрения: Газпром может не только обеспечивать четверть бюджета, но и стать локомотивом реформ ..., заявил на собрании акционеров Рем Вяхирев.

Заголовок текста: который соответствует формулировке основной темы документа, содержит три основных понятия: *Газпром*, *финансы* и *государство*. Следовательно, подтемы текста должны раскрывать отношения между этими тремя понятиями. Рассмотрим, какие же слова и словосочетания входят в тематические узлы основных понятий, то есть используются в качестве ссылок на эти понятия в тексте.

<i>Газпром</i>	<i>финансы</i>	<i>государство</i>
<i>Концерн</i>	<i>федеральный бюджет</i>	<i>федеральный бюджет</i>
<i>компания</i>	<i>налоговый доход</i>	<i>бюджет</i>
<i>общество</i>	<i>платеж</i>	<i>пенсионный фонд</i>
<i>акционер</i>	<i>выплата</i>	<i>казначейство</i>
<i>акция</i>	<i>задолженность</i>	<i>бюджетные организации</i>
<i>газ</i>	<i>банк</i>	<i>минобороны</i>
<i>Вяхирев</i>	<i>бюджет</i>	<i>части МВД</i>
	<i>пенсионный фонд</i>	<i>академия наук</i>
	<i>прибыль</i>	
	<i>акция</i>	

Собственно расположение в тексте слов, соответствующих этим тематическим узлам, и создает эффект лексических цепочек.

4. Структура текста и имплицитные концептуальные связи

Рассматриваемый текст представляет собой один из примеров таких текстов, в которых структурная организация текста строится, в большей степени, на относительно скрытых лексических и энциклопедических связях между словами (более двадцати), чем на использовании местоимений (5 указательных и личных местоимений) и связочных слов.

Кроме того, в данном тексте исключительную важность имеют лексические связи достаточно сложной природы, в отличие от относительно более изученных в качестве механизма связности текста лексических повторов, родовых и видовых связей. Рассмотрим на примере этого текста основные типы понятийных отношений, которые могут использоваться для организации текста. Классификация отношений будет производиться в зависимости от типа пути, которым можно соединить эти два понятия в понятийной системе Тезауруса:

1) Лексическая связность осуществляется на основе одного и того же концепта: реализуется лексическими повторами или синонимами: *федеральный бюджет* - *госбюджет*.

2) Лексическая связность осуществляется на основе понятий, которые непосредственно связаны между собой: *Газпром* - *ВЫШЕ- компания*, *акция* - *АССОЦИАЦИЯ* - *акционер*.

3) Путь от одного понятия к другому состоит из совокупности связей ВЫШЕ и ЦЕЛОЕ, например:

- *акция* - ВЫШЕ - *ценная бумага* - ЦЕЛОЕ - *фондовый рынок* - ВЫШЕ - *рынок* - ЦЕЛОЕ - *финансы*.
- *казначейство* - ЦЕЛОЕ - *министерство финансов* - ВЫШЕ - *министерство* - ВЫШЕ - *орган исполнительной власти* - ЦЕЛОЕ - *государственная власть* - ЦЕЛОЕ - *государство*.
- *министерство обороны* - ВЫШЕ - *министерство* - ВЫШЕ - *бюджетная организация*

Вывод таких лексических связей базируется на свойствах транзитивности отношений ВЫШЕ-НИЖЕ и ЧАСТЬ-ЦЕЛОЕ.

4) Путь состоит из двух однонаправленных фрагментов: один фрагмент пути - совокупность отношений ВЫШЕ и ЦЕЛОЕ, затем перегиб и далее совокупность отношений ЧАСТЬ и НИЖЕ, то есть имеет место наследование элементами одной ветви свойств элементов другой ветви, например:

- *Газпром* - ВЫШЕ - *акционерное общество* - ЧАСТЬ - *акционер*;

5) Возможен и перегиб пути другого типа: сначала отношения НИЖЕ или ЧАСТЬ, а затем отношение ЦЕЛОЕ или ВЫШЕ, например:

- *платеж* - НИЖЕ - *безналичный платеж* - ВЫШЕ - *банковская операция* - ЦЕЛОЕ - *банковская деятельность* - АСЦ₁ - *банк*.
- *пенсионный фонд* - НИЖЕ - *государственный пенсионный фонд* - ЦЕЛОЕ - *пенсионное обеспечение* - ВЫШЕ - *социальная защита* - ЦЕЛОЕ - *социальная сфера* - ЦЕЛОЕ - *государство*

5. Выявление тематических узлов в тексте

Анализируя приведенные тематические узлы, видно, что выявление тематических узлов и, следовательно, основных понятий в тексте представляет собой сложный процесс. Эта сложность связана со следующими проблемами:

1) необходимость включения в формулировку темы таких обобщающих понятий как *финансы* и *государство* не следует сразу же из первого абзаца текста (для первого абзаца достаточно понятий *выплата и госбюджет*), а вытекает из анализа текста в целом, из необходимости поиска связей между различными абзацами текста;

2) для многих из упоминаемых в тексте понятий для организации текста используется сразу или последовательно несколько типов связей. Так, для понятия *Газпром* - это связи с понятиями *концерн, акционерное общество, компания, газ*; понятие *госбюджет* одновременно связано с двумя важными для текста понятиями - *финансы и государство* и т.д.;

3) понятия основной темы связаны между собой концептуальными путями того же типа, что использовались для построения самих тематических узлов, а именно:

понятия *Газпром* и *государство* связаны путем типа 3) (см.п.4):

государство -НИЖЕ -Россия - ЧАСТЬ - Российская компания - НИЖЕ - Газпром;

понятия *финансы и государство* связаны между собой путем типа 5) (см.п.4):

финансы - ЧАСТЬ - бюджет -НИЖЕ - госбюджет -ЦЕЛОЕ - государство.

4) из п.4 возникает закономерный вопрос, насколько использование одного и то же отношения или одного и того же пути между одними и теми же понятиями может варьироваться: в одном тексте эти понятия являются элементами одного и того же тематического узла, в другом тексте вошли в разные тематические узлы.

Таким образом, правильный вывод тематических узлов текста, а значит и его основных понятий требует многоаспектного анализа текста. Даже человек не может определить всю совокупность лексических связей, необходимых для организации данного текста, двигаясь от предложения, к предложению, не прочитав весь текст целиком. Для компьютерной системы это не означает невозможность автоматического выявления основных понятий текста, но только то, что для такого анализа сначала необходимо восстановить понятийную сеть, стоящую за данным текстом, выявить потенциальные связи, которые могут быть использованы в тематических узлах, а затем анализировать их распределение в предложениях текста.

6. Вариативность использования концептуальных связей в основных тематических узлах

В предыдущих параграфах мы выделили три множества концептов, явно или неявно затронутых в некотором тексте: MC_1 - множество понятий основной темы текста, MC_2 - все остальные понятия текста, подмножество которых MC_3 , используется для ссылки на MC_1 . Пусть C_1 - понятие из MC_1 , а понятие $C_{1,i}$ принадлежит его тематическому узлу, т.е. входит в подмножество MC_3 , тогда отношение $C_1 - C_{1,i}$ используется как инструмент лексической связности текста. Если имеется текст, в котором понятия C_1 и $C_{1,i}$ оба находятся в основной теме текста, то отношение $C_1 - C_{1,i}$ меняет свой статус, и: следовательно, C_1 и $C_{1,i}$ не могут находиться в одном тематическом узле. Здесь и кроются причины вариативности использования концептуальных отношений в качестве инструмента тематической организации текста. Рассмотрим примеры.

Как известно, отношение ЧАСТЬ-ЦЕЛОЕ часто используется для организации связности между предложениями текста. Такими частями, например, для Российской Федерации являются ее административные единицы: области, края и т.д. Знание о принадлежности той или иной области России многократно используется для организации текстов, например,

Пример 5а.

Стороны договорились о визите в ближайшее время в Россию министра иностранных дел Ирана. Была там тогда достигнута и договоренность о передаче гуманитарной помощи вынужденным переселенцам из Чечни.

Однако статус Чечни стал изменяться, и в текстах стала видна тенденция рассматривать Чечню как отдельную сущность

Пример 5б.

Проведен опрос 185 воронежцев. ... Были заданы три вопроса: 1. Считаете ли вы Чечню территорией России? .. 46,48 процента опрошенных считают Чечню территорией России. И ровно столько же ее не считают территорией РФ... О том, что выход Чечни из состава России может послужить началом развала Федерации, никто не задумывается.

Основную тему данного текста можно сформулировать следующим образом: *социологический опрос воронежцев, посвященный статусу Чечни в составе России.*

Рассмотрим другой пример. Если мы посмотрим, на абзацы 2 и 3 примера из п.3 мы увидим, что для связности текста используется знание о том, что Газпром является акционерным обществом, имеет акции. То есть можно сказать, что текст был о Газпроме и в частности обсуждались выплаты по акциям. Рассмотрим однако пример текста, основанный на похожих концептуальных знаниях:

Пример 5в.

ИНОСТРАНЦЫ ПОДХЛЕСТНУТ СПРОС НА АКЦИИ СБЕРЕГАТЕЛЬНОГО БАНКА РОССИИ

СБЕРБАНК России принял решение ввести квоты на покупку акций нерезидентами. Суммарная доля уставного капитала, которая может принадлежать иностранным инвесторам, установлена в размере пяти процентов, при этом один акционер имеет право приобрести не более одного процента бумаг... Учитывая сказанное, можно предположить, что спрос на акции Сбербанка со стороны как отечественных, так и иностранных Сбербанк со стороны как отечественных, так и иностранных инвесторов сохранится на прежнем, высоком уровне, а их курсовая стоимость продолжит свой рост.

Пересказывая содержание этого текста мы уже не можем сказать, что текст был о Сбербанке и о стоимости акций Сбербанка, поскольку само такое конкретное понятие как *акция Сбербанка* и являлось фокусом обсуждения текста, т.е. и понятие *акция*, и понятие *Сбербанк* являются основными понятиями для данного текста.

7. Лексическая связность vs. совместная встречаемость в предложениях текста

Казалось бы, что в результате предшествующих рассмотрений мы попали в замкнутый круг. Для того, чтобы найти основные понятия текста, нужно найти наиболее “сильные” лексические цепочки в тексте, но только основная тема текста определяет, каково будет разбиение понятийной сети, стоящей за текстом, на такие цепочки.

Выходом здесь является следующее соображение. Текст посвящен не только и не столько обсуждению своих основных понятий, сколько обсуждению отношений между этими основными понятиями. Поэтому, не только основные понятия и элементы их тематических узлов наиболее частотны в тексте, но и пары различных основных понятий и элементов их тематических узлов должны встречаться в предложениях текста чаще, чем, например, элементы одного и того же тематического узла. Отсюда следует, что если два понятия, связаны по Тезаурусу, одним из типов концептуальных путей 1) - 5), однако относительно часто встречаются в одних и тех же предложениях текста, то вряд ли одно из них входит в тематический узел другого.

Действительно, рассмотрим, каким именно образом распределены в предложениях текста примера пары понятий: 1) *Газпром (компания, общество, концерн) - госбюджет (бюджет)*, 2) *Газпром (компания, общество) - государство*, 3) *госбюджет (бюджет) - государство*. Мы видим, что первые пары употреблялись в одних и тех же предложениях - 5 раз, вторые пары - 3 раза, третьи пары - один раз в разных частях сложного предложения. Также часто встречаются в одних и тех же предложениях примера 5б понятия *Россия* и *Чечня*, и понятия *акции* и *Сбербанк* в предложениях примера 5в.

Заключение

Все вышесказанное позволяет нам сделать следующие основные выводы.

1) обсуждаемые в тексте понятия связаны между собой разнообразными семантическими и энциклопедическими связями, многие из которых могут использоваться для организации тематической структуры того или иного текста;

2) невозможно, двигаясь по порядку следования предложений в тексте, с достаточной надежностью находить совокупности слов, соответствующие элементам основной темы текста, поскольку точный набор задействованных связей, само разбиение всего множества слов текста на совокупности близких слов в значительной мере зависит от изложения текста в целом, от основной темы документа;

3) для того чтобы автоматически выявлять основные понятия текста на основе имеющегося лингвистического ресурса необходимо, как минимум:

3.1) восстановить сеть понятий, используемых в данном тексте;

3.2) разбить сеть на совокупность тематических узлов, используя знания о способах соединения понятий внутри тематических узлов, учитывая информацию о взаимной встречаемости связанных по Тезаурусу понятий в одних и тех же предложениях данного текста;

3.3) среди всей совокупности тематических узлов выделить основные тематические узлы.

Литература

ван Дейк Т.А., Кинч В. 1988. Стратегии понимания связного текста. // Новое в зарубежной лингвистике. Вып. 23. - М.: Прогресс. - С.153-211.

Крылов С. 1984. Детерминация имени в русском языке: теоретические проблемы // Семиотика и информатика. - 1984. - Вып.23. - С.124 -154.

Лукашевич Н.В., Салий А.Д. 1997. Представление знаний в системе автоматической обработки текстов // НТИ. Сер.2. - 1997 - N3.

Лукашевич Н.В. 1999. От общеполитического тезауруса к тезаурусу русского языка в контексте автоматической обработки больших массивов текстов. - Труды международного семинара Диалог-99, т.2, с.184 -190.

Селезнев 1987. Референция и номинация. - Моделирование языковой деятельности в интеллектуальных системах. - М. Наука. - с. 64-77.

Barzilay R., Elhadad M. 1997. Using Lexical Chains for Text Summarization. - ACL/EACL Workshop Intelligent Scalable Text Summarization.- Madrid.

Halliday M., Hasan R. (1976) : *Cohesion in English*. Longman, London.

Loukachevitch N.V., Salii A.D., Dobrov B.V. 1999. Thesaurus for Automatic Indexing: Structure, Developement, Use. In Proceedings of International Congress "Terminology and Knowledge Engineering". p. 343-355.

Morris J., Hirst G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of a text. *Computational Linguistics*, 17 (1), 21-48.