УДК 81'33
ББК Ш11 *ГСНТИ 16.31.21* *Код ВАК 10.02.19*

А. В. Добров
Санкт-Петербург, Россия

КОМПЛЕКСНЫЙ ЛИНГВИСТИЧЕСКИЙ ПОДХОД К АВТОМАТИЧЕСКОЙ РУБРИКАЦИИ НОВОСТНЫХ СООБЩЕНИЙ

Аннотация. *Непрерывный рост объемов информации в средствах массовой коммуникации приводит к повышению трудоемкости ручной классификации текстов. Для автоматизации этой деятельности создан ряд компьютерных средств, но уровень их эффективности недостаточно высок для того, чтобы упростить деятельность экспертов. Предлагается подход к созданию систем автоматической рубрикации новостных сообщений, основанный на компьютерных методах комплексного лингвистического анализа текстов, анализируются способы оценки эффективности таких систем.*

Ключевые слова: *новостные сообщения; медиадискурс; автоматическая рубрикация текстов; тематическая структура дискурса; автоматический семантический анализ.*

Сведения об авторе: *Добров Алексей Владимирович, аспирант кафедры математической лингвистики филологического факультета.*

Место работы: Санкт-Петербургский государственный университет; ОАО «Линукс Инк».

Контактная информация: *199034, г. Санкт-Петербург, Университетская наб. д. 7.*
e-mail: a.dobrov@linux-ink.ru.

A. V. Dobrov
St. Petersburg, Russia

A COMPLEX LINGUISTIC APPROACH TO AUTOMATIC CLASSIFICATION OF NEWS REPORTS

Abstract. *The continuous growth of the amount of information in mass media leads to the increase of labour-intensity of manual text classification. A set of computer tools is developed to automate this process, but the level of effectiveness of these tools is not high enough to simplify the work of the experts. This article introduces an approach to development of automatic classification systems of news reports, based on machine-driven complex linguistic analysis. Some techniques of evaluation of effectiveness of those systems are discussed.*

Key words: *news reports; media discourse; automatic text classification; topical structure of discourse; automatic semantic analysis.*

About the author: *Dobrov Alexey Vladimirovich, Post-graduate Student of the Chair of Computational Linguistics of the Philological Faculty.*

Place of employment: the St.Petersburg State University; OJSC "Linux Ink".

Новостные сообщения, или сообщения информационных агентств (напр., РИА Новости, ИТАР-ТАСС, Рейтер и т. п.), — это одна из разновидностей новостных текстов, нуждающихся в эффективной и быстрой рубрикации. Новостные сообщения появляются каждый час и распределяются по сводным и тематически ориентированным новостным потокам. Скорость и качество такого распределения зависят от деятельности коллективов экспертов, осуществляющих рубрикацию новостных сообщений вручную.

Рост объемов информации в средствах массовой коммуникации, и в частности в сети Интернет, приводит к постоянному повышению трудоемкости этой деятельности и к необходимости в совершенствовании качества и скорости поиска информации и организации ее хранения. Для решения этих задач создается компьютерный инструментарий, позволяющий в той или иной мере автоматизировать и упростить процесс рубрикации текстов, однако уровень эффективности систем автоматической рубрикации текстов недостаточно высок для того, чтобы упростить деятельность экспертов.

В отличие от других видов новостных текстов, новостные сообщения выполняют в первую очередь информационную функцию, а не функцию воздействия, поэтому в наименьшей мере проявляют те особенности публицистического функционального стиля, которые затрудняют автоматический лингвистический анализ. В новостных сообщениях используется сравнительно меньшее количество средств художественной выразительности, подавляющее количество предложений — повествовательные, либо констатирующие тот или иной факт, либо передающие чей-либо комментарий к этому факту в виде прямой речи. Как и другие новостные тексты, новостные сообщения характеризуются четкой тематической структурой дискурса. В отличие от информационной аналитики и комментария, многих видов публицистики, тематических материалов группы «features» (по классификации Т. Г. Добросклонской [см. Добросклонская 2005]) и рекламы, новостные сообщения характеризуются также относительной краткостью и единообразием структуры текста и структур составляющих предложений. Поэтому при оценке эффективности систем автоматической рубрикации текстов новостные сообщения представляют собой одну из разновидностей материала, на котором статистические методы оценки эффективности дают наиболее достоверные результаты. К таким разновидностям языкового материала традиционно относятся также нормативно-правовые документы, в отличие от которых новостные сообщения

© Добров А. В., 2011

характеризуются меньшим количеством специальных терминов и сложных юридических формулировок, что упрощает применение лингвистических средств компьютерного анализа. Этими особенностями новостных сообщений обусловлена также бóльшая сложность применения статистических методов для их автоматической рубрикации: благодаря значительному количеству общеупотребительной лексики термины, характерные для конкретных рубрик, оказываются менее частотными, а отсутствие сложных типовых формулировок не дает возможности производить уточнение алгоритмов выбора рубрик путем их ориентации на конкретные цепочки формальных оболочек словоформ.

Поэтому наряду с коллекциями правовых документов многие текстовые коллекции, предназначенные для оценки эффективности работы систем автоматической рубрикации текстов, состоят из новостных сообщений, например коллекция «Reuters-21578» [Lewis 2004].

Задача автоматической рубрикации новостных сообщений может пониматься по-разному, так как устоявшегося определения термина «автоматическая рубрикация» в современной лингвистике не существует. Тем не менее в различных публикациях приводятся некоторые определения, из которых наиболее абстрактной представляется формулировка М. С. Агеева, Б. В. Доброва и Н. В. Лукашевич: «...отнесение порции информации к одной или нескольким категориям из ограниченного множества» [Агеев, Добров, Лукашевич 2008: 25].

При автоматической рубрикации тексты распределяются по заранее специфицированным рубрикам. В зависимости от того, каким образом специфицируются эти рубрики, можно выделить два класса методов организации систем автоматической рубрикации текстов: методы, основанные на знаниях, и методы, основанные на машинном обучении, при применении которых используется коллекция документов, предварительно распределенных по рубрикам вручную. Принципиальное различие между этими двумя подходами, как представляется, состоит не в технологии автоматической рубрикации, а в методике предварительного получения описаний рубрик: при «инженерном» подходе эти описания строит человек, а при применении алгоритмов машинного обучения — машина. По данным исследования М. С. Агеева и др., «... системы рубрикации, основанные на машинном обучении, имеют серьезные проблемы даже на относительно простом рубрикаторе: 50 % F-меры означает, что только половина документов получила правильные рубрики» [Там же: 27].

Под описанием рубрик при инженерном подходе к автоматической рубрикации авторы понимают «некоторое выражение на основе слов и (или) терминов реальных текстов» [Там же: 31], а не описание понятий, стоящих за различными значениями этих слов. Такой подход приводит к ряду затруднений, на которые указывают многие авторы, обусловленных многозначностью слов (слова в тексте могут употребляться не в тех значениях, на которые рассчитывал эксперт) и проблемой ложной корреляции. Ложная корреляция возникает в случаях, когда правило отнесения текста к рубрике опирается на присутствие в нем двух не связанных друг с другом слов. Например, если для отнесения к рубрике «*Экономические реформы*» считать достаточным наличие в тексте слов «*экономический*» и «*реформа*», то ложная корреляция может произойти, если в тексте шла речь о судебной реформе и были упомянуты экономические вопросы. Причина возникновения проблемы «ложной корреляции» состоит в не вполне корректном подходе к созданию лингвистического обеспечения системы авторубрикации: для отнесения текста к рубрике «*Экономические реформы*» необходимо присутствие в тексте не слов «*экономический*» и «*реформа*», а темы ‘экономические реформы’, представленной относящимися к ней понятиями, выраженными синтаксически и семантически связанными языковыми единицами. Для этого, однако, описывать рубрики необходимо в форме лингвистических моделей, отражающих связи между понятиями и выражающими их единицами — синтаксическими структурами, обладающими сложной семантикой, позволяющей относить или не относить понятия, а не отдельные слова, к той или иной «рубрике».

Таким образом, современные системы автоматической рубрикации текстов характеризуются рядом недостатков, обусловленных тем, что в них не задействован синтаксический и семантический анализы текстов. Основные недостатки систем автоматической рубрикации текстов можно было бы ликвидировать, если бы обрабатываемыми единицами были не отдельно взятые словоформы или сопоставляемые им версии лемматизации с неразрешенной неоднозначностью, а элементы полноценного семантического представления обрабатываемого текста.

При использовании комплексного лингвистического анализа текста образы рубрик могли бы представлять собой формулы, сходные по структуре с теми, которые применяются в современных системах автоматической рубрикации текстов, но в функции переменных в этих формулах выступали бы не лексические единицы, а концептуальные структуры, соответствующие их значениям (ниже будут даны некоторые пояснения, относящиеся к способам организации и методам автоматического выявления этих структур). Например, образ рубрики «*Экономическая реформа*» включал бы в себя концепт ‘экономическая реформа’, а не формулу вида «*(экономический ИЛИ экономика) И реформа*». Такой подход решил бы проблему ложной корреляции: текст, в содержании кото-

рого концепты ‘экономика’ и ‘реформа’ не связаны друг с другом, не мог бы быть отнесен к рубрике «*Экономическая реформа*».

Комплексный лингвистический анализ текста дает также возможность учета его тематической (коммуникативной) структуры. Исходными данными для выявления тематической структуры дискурса является информация о когезии и об актуальном членении. Как отмечают различные авторы, на основании этих данных может быть выявлена иерархическая (древовидная) структура тематической прогрессии текста (см., напр., тематические прогрессии Данеша — [Danes 1974]). При учете тематической структуры дискурса основанием для ранжирования рубрик становится не только частотность соответствующих им концептуальных структур, но и их место в структуре тематической прогрессии.

Различные подходы к описанию тематической структуры дискурса подробно рассматриваются в работе [Филиппов 2003]. Основным недостатком существующих подходов к моделированию тематической структуры дискурса представляется то, что в подавляющем большинстве эти подходы основаны на идее единственности этой структуры. Предполагается, что в каждом предложении можно четко выделить часть, относящуюся к теме, и часть, относящуюся к реме, при этом тема любого не первого предложения текста непременно является ремой одного из предшествующих ему предложений. На чем основаны эти предположения, остается неясным. В структуре тематической или рематической составляющих каждого предложения может быть любая составляющая, причем в составе сложной рематической составляющей могут присутствовать тематические компоненты. Тематические компоненты предложений в составе текста не обязательно одновременно являются рематическими компонентами предшествующих им предложений — они могут относиться к любому компоненту коммуникативной ситуации, включая общие для говорящего и слушающего сведения об обсуждаемой действительности. Наконец, в одном предложении могут фигурировать тематические компоненты, являющиеся одновременно рематическими в нескольких предшествующих предложениях, что дает основания предполагать возможность сосуществования нескольких тематических прогрессий в рамках одного текста или, что представляется в большей степени целесообразным, представлять тематическую прогрессию в виде сетевой, а не древовидной структуры. Следует отметить, что указанные явления встречаются в подавляющем большинстве текстов, поэтому в качестве примера можно было бы взять любое новостное сообщение. Рассмотрим, например, тематическую прогрессию в сообщении агентства ИТАР-ТАСС от 29 июня 2011 г., опубликованном в электронной форме в сети Интернет по адресу http://www.itar-tass.com/c1/175781.html (для краткости возьмем только заголовок и первый абзац).

Первое предложение — заголовок: *Греческие демонстранты на площади у парламента забрасывают полицию бутылками с зажигательной смесью*.

Второе предложение (первое в основном тексте сообщения) связано с заголовком одновременно четырьмя анафорическими связями: *Демонстранты начали забрасывать полицию бутылками с зажигательной смесью и камнями у парламента Греции в рамках акции протеста против ожидаемого принятия сегодня законодателями программы жесткой экономии до 2015 года*.

Ниже приводится таблица с указанием анафорических связей и статуса синтаксических составляющих (T — тема, R — рема) в актуальном членении предложения.

Предложение1	**Предложение2**
Греческие демонстранты (T)	Демонстранты (T)
забрасывают полицию (R)	начали (R) забрасывать полицию (T)
бутылками с зажигательной смесью (R)	бутылками с зажигательной смесью (T)
у парламента (R).	у парламента Греции (T).

Третье предложение связано со вторым двумя анафорическими связями, при этом одна из них — эллиптическая (эллиптированный фрагмент приведен в угловых скобках): *В ответ спецназовцы в касках и противогазах, прикрывающиеся пластиковыми щитами, распыляют в толпу слезоточивый газ.*

Предложение2	**Предложение3**
начали (R) забрасывать полицию (T)	В ответ (R) <на это> (T)
Демонстранты (T)	в толпу (T)

Четвертое предложение является сложносочиненным, его компоненты целесообразно рассматривать по отдельности, так как между ними наблюдается анафорическая связь: *В ходе ожесточенных стычек манифестанты бьют полицейских палками от транспаран-*

тов, в ответ спецназовцы пускают в ход резиновые дубинки.

Первая часть четвертого предложения связана со вторым предложением двумя анафорическими связями:

Предложение3	**Предложение4.1**
Демонстранты (T)	манифестанты (T)
полицию (T)	Полицейских (T)

Вторая часть четвертого предложения связана одной анафорической связью с первой частью и одной анафорической связью с третьим предложением.

Предложение4.1	**Предложение4.2**
бьют полицейских палками от транспарантов (R)	В ответ (R) <на это> (T)
Предложение3	**предложение4.2**
спецназовцы (R)	Спецназовцы (T)

Пятое предложение не содержит в себе прямых анафорических связей ни с одним из предыдущих: *Есть раненые, задержаны несколько человек.* В этом предложении описывается итог тех столкновений, о которых шла речь во всех предыдущих предложениях, поэтому можно предположить, что в нем опущен компонент <в результате этих столкновений>, тогда через него пятое предложение анафорически связано со всеми предыдущими.

В первом абзаце большинство предложений связано с предыдущими предложениями несколькими анафорическими связями, причем некоторые предложения связаны одновременно с несколькими предшествующими, а одно предложение (четвертое) связано само с собой. Большинство анафорических связей выражено языковыми единицами, находящимися в отношениях синонимии (*демонстранты — манифестанты*) или гиперонимии (*демонстранты — толпа*), многие связи выявляются за счет восстановления эллиптированных компонентов. Подобная система анафорических связей показывает, что элементами тематической прогрессии целесообразно считать не предложения, а концептуальные структуры, соответствующие их составляющим.

Такой подход к моделированию тематической структуры дискурса дает возможность формализации методики ее выявления на основании сопоставления концептуальных структур, соответствующих синтаксическим составляющим: для каждой пары таких структур предполагается анафорическая связь, если второй элемент пары является тематическим, и между корневыми элементами этих структур существуют отношения эквивалентности (синонимии) или наследования (гиперонимии). Тематическая структура дискурса формируется в результате выявления анафорических связей в виде сетевой структуры (графа), узлами которой являются концептуальные структуры, а дугами — анафорические связи. Элементы концептуальных структур предложений, не входящие в анафорические связи, остаются связанными с анафорическими компонентами за счет семантико-синтаксических связей. На основании такого представления тематической структуры дискурса можно вычислить ранг каждой семантической структуры, представленной в тексте, в соответствии со следующими правилами:

- ранг концептуальной структуры, выраженной синтаксической составляющей, не связанной ни напрямую, ни опосредованно никакими анафорическими связями с другими составляющими, равен единице;
- концептуальные структуры, выраженные тематическими составляющими, связанными с другими составляющими напрямую или опосредованно анафорическими связями, обладают одинаковым рангом;
- концептуальные структуры, выраженные рематическими составляющими, связанными с другими составляющими опосредованно анафорическими связями, обладают рангом на единицу большим, чем ранг антецедента.

В соответствии с этими правилами в приведенном отрывке рангом, равным единице, обладают, например, структуры ‘греческие демонстранты’, ‘полицейские’, ‘забрасывать полицию’, ‘бутылки с зажигательной смесью’, ‘бить полицейских палками от транспарантов’. Рангом, равным двум, обладают, например, структуры ‘акция протеста’, ‘камни’, ‘принятие программы жесткой экономии’, ‘спецназовцы’. Рангом 3 обладают, например, структуры ‘каски’, ‘противогазы’, ‘резиновые дубинки’.

Организация концептуальных структур рассматривается в научной литературе по-разному. Термин «концептуальный граф» был впервые введен Дж. Ф. Сова в работе [Sowa 1976]. В первоначальном понимании концептуальный граф — это двудольный направленный граф, состоящий из двух типов узлов: концептов и концептуальных отношений, или просто отношений.

В ряде работ по синтаксической семантике формализм концептуальных графов использу-

ется для моделирования семантики предложения [см., напр.: Богатырев, Тюхтин 2009; Палагин, Кривой, Петренко 2009 и т. п.]. Далеко не всегда сетевое представление семантики предложения в виде концептуального графа соответствует определению Дж. Совы: часто отношения между концептами моделируются не при помощи узлов концептуального графа, а при помощи его дуг. Такой граф, в котором отношения между концептами моделируются при помощи дуг, иногда называют семантической сетью: «Семантические сети представляются в виде направленного графа, вершины которого соответствуют объектам (понятиям, сущностям) предметной области, а дуги — отношениям (связям) между объектами» [Палагин, Кривой, Петренко 2009: 77]. В соответствии с этим определением, частным случаем семантической сети можно считать семантическое представление, предложенное еще И. А. Мельчуком в модели «Смысл <-> Текст» [см.Мельчук 1974].

При обоих подходах трудности возникают при моделировании отношений более чем первого порядка, т. е. таких, которые устанавливаются между другими отношениями. Например, союз «и», связывающий два предиката, выражает отношение между двумя отношениями. В модели Дж. Совы такая возможность вообще не предусмотрена, ведь как только в концептуальном графе появится дуга, связывающая два отношения, будет нарушено требование двудольности. В модели семантической сети видится две возможности решения этой задачи: можно постулировать возможность для некоторых дуг связывать между собой не узлы, а дуги, или интерпретировать узлы и дуги с одинаковыми метками как одни и те же концепты. В первом случае концептуальная структура по определению перестанет быть графом, что приведет к значительным затруднениям при ее автоматической обработке. Во втором случае в рамках концептуальной структуры не будут явным образом разграничены концепты и отношения. В рамках предлагаемого в данной статье подхода предпочтение отдается второму решению, так как, во-первых, концептуальные отношения можно рассматривать как частный случай концепта, а во-вторых, сохранение структуры графа позволяет применять к концептуальной структуре существующие алгоритмы обработки графов.

Предлагаемая модель организации концептуальной структуры может быть проиллюстрирована на примере первого предложения рассматривавшегося выше отрывка (рис. 1). Этот граф был построен в результате обработки синтаксической структуры предложения, полученной в ходе синтаксического анализа (рис. 2). Следует отметить, что наименования синтаксических составляющих и их вложенность были получены на основании разрабатываемой автором данной статьи формальной модели русского синтаксиса.

Построение концептуальной структуры на основе дерева синтаксических составляющих было выполнено в соответствии со следующими принципами:

1) концептуальная структура терминальной синтаксической составляющей представляет собой одно из значений соответствующей этой составляющей словоформы;

2) концептуальная структура сложной идиоматической синтаксической составляющей представляет собой одно из значений соответствующей этой составляющей идиомы;

3) концептуальная структура сложной неидиоматической синтаксической составляющей вычисляется из концептуальных структур ее дочерних составляющих при помощи функции, соответствующей типу родительской составляющей;

4) выбор того или иного значения в пп. 1, 2 обусловлен возможностью вычисления функции в п. 3

В случае с сигнификативной именной группой *Греческие демонстранты* выбор значения производить не приходится: оба ее компонента — словоформы однозначных лексических единиц. Соответственно, вычисление концептуальной структуры этой составляющей сводится к включению в сигнификат понятия ‘демонстрант’ признака ‘относящийся к Греции’, обозначаемого прилагательным *греческий*. В результате связывания понятия ‘демонстрант’ с грамматическим значением множественного числа значение словоформы *демонстранты* интерпретируется при составлении концептуального графа как ‘группа (совокупность), включающая в себя элементы класса «демонстрант»’.

Проблема выбора значения возникает при интерпретации группы переходного глагола *забрасывают полицию бутылками с зажигательной смесью*. В этом предложении глагол *забрасывать* употреблен в значении ‘бросать в большом количестве, ошеломляя’, а существительное *полиция* — в значении ‘совокупность полицейских’. Оба значения не являются основными для этих лексических единиц. Глагол «забрасывать» (НСВ к «забросать») омонимичен глаголу «забрасывать» (НСВ к «забросить»). В данном случае выбор первого варианта основан на несоответствии рамки валентностей глагола «забросить», не содержащей в себе инструмента, фактически наблюдаемому инструменту, выраженному пассивным субъектом «бутылками с зажигательной смесью». В результате этого несоответствия не может быть вычислена функция заполнения рамки валентностей, соответствующая группе переходного глагола. Выбор значения ‘совокупность полицейских’ обусловлен семантическим ограничением на пациенс, состоящим в том, что в роли пациенса в данном случае может выступать только человек. Это ограничение возникает благодаря семе ‘ошеломляя’, представленной в значении глагола «забрасывать».

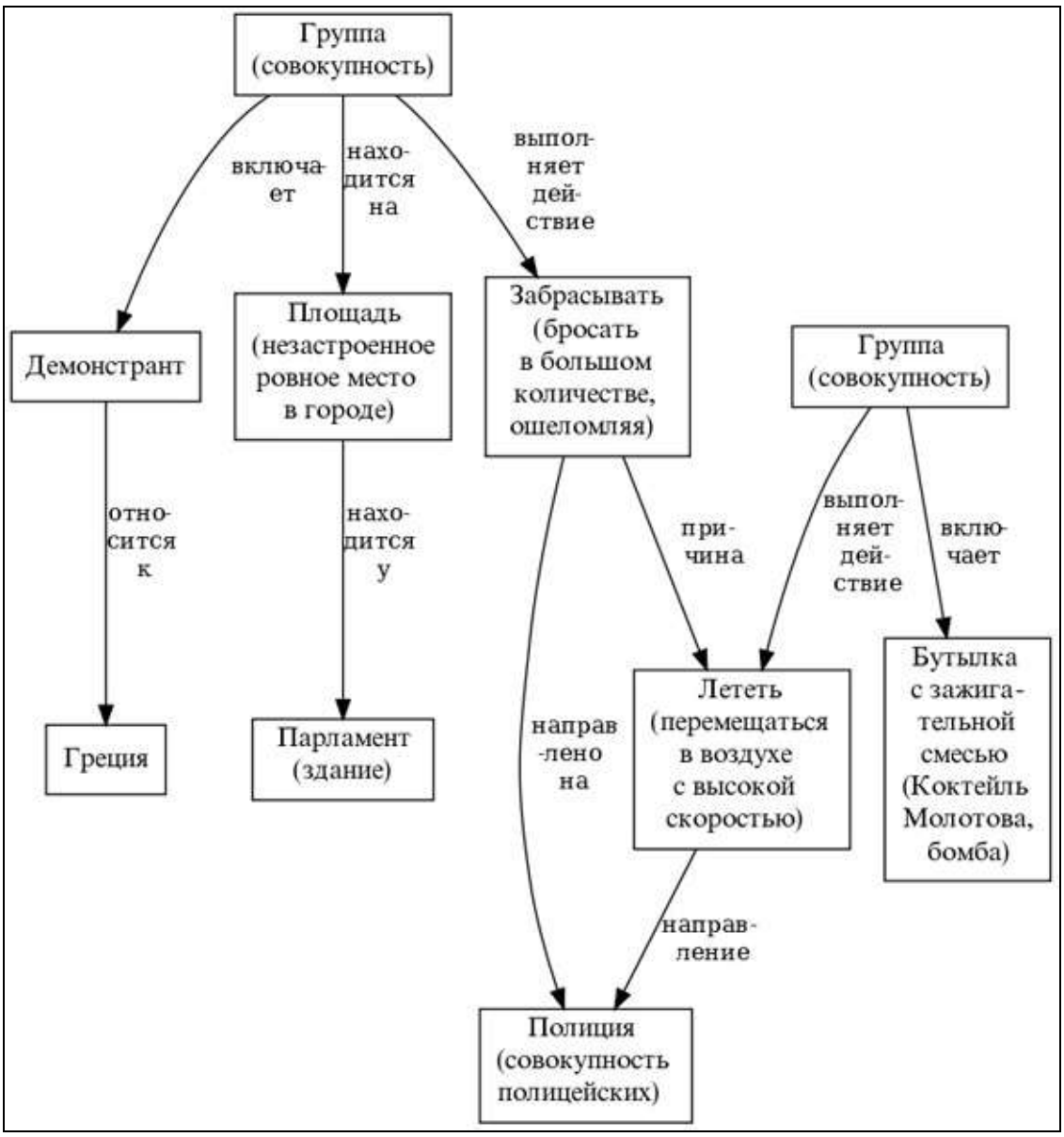

Рисунок 1

Примером действия принципа 2 может служить сигнификативная именная группа *бутылками с зажигательной смесью*, представляющая собой идиоматическое выражение, значение которого — ‘Коктейль Молотова (бомба)’ — невозможно вывести из его частей.

Использование концептуальных структур, вычисляемых из синтаксических в соответствии с указанными принципами, для выявления тематической структуры новостного сообщения и его автоматической рубрикации решает проблему лексической неоднозначности и проблему ложной корреляции. Для этого необходимо использовать компьютерную модель лексики и грамматики, содержащую в себе информацию о синтаксических составляющих, функциях, позволяющих вычислять их значения, а также о значениях лексических единиц, синонимических и гиперонимических отношениях между ними и о соответствующих этим значениям рамках валентностей.

Использование комплексного лингвистического анализа текстов для их автоматической рубрикации на основании формальных семантико-синтаксических моделей может существенно повысить эффективность работы систем автоматической рубрикации текстов. Чтобы установить доподлинно, в какой степени применение комплексного лингвистического анализа повышает эффективность систем автоматической рубрикации текстов, необходимо создать подобную систему и произвести экспериментальное исследование с целью сравнения эффективности работы существующих систем автоматической рубрикации текстов с эффективностью работы созданной системы.

Оценка эффективности работы системы автоматической рубрикации текстов производится путем сравнения результатов ее работы с эталонной рубрикацией набора документов. В качестве эталона используется коллекция документов, отрубрицированных вручную. Такой подход может вызвать некоторые сложности, так как эффективность ручной рубрикации зависит от ряда факторов. Эксперты, производящие ручную рубрикацию, могут работать не

вполне последовательно, что может отразиться на качестве их работы.

Работа одного эксперта вряд ли может считаться эталоном при оценке качества работы системы автоматической рубрикации. Как отмечалось в работе [Добров 2010], релевантность той или иной рубрики документу представляет собой шкалируемую величину, поэтому эталонная рубрикация должна содержать в себе оценку этой величины для каждого отнесения документа к рубрике. Тогда эталонная рубрикация одной и той же коллекции документов может быть произведена несколькими экспертами. При этом для каждого отнесения документа к рубрике можно произвести расчет математического ожидания и дисперсии значения релевантности, что позволяет, в соответствии с Т-критерием Стьюдента, оценить статистическую значимость различий между результатами работы системы автоматической рубрикации текстов и эталонной рубрикацией.

Для оценки эффективности систем автоматической рубрикации текстов принято использовать метрики, сходные с теми, которые используются для оценки эффективности работы информационно-поисковых систем.

Меры точности и полноты были введены и описаны в 1955 году А. Кентом и его коллегами, разработавшими «систему оценки» информационно-поисковых систем, включающую в себя методы статистической выборки для оценки числа не найденных релевантных документов [Kent et al. 1955].

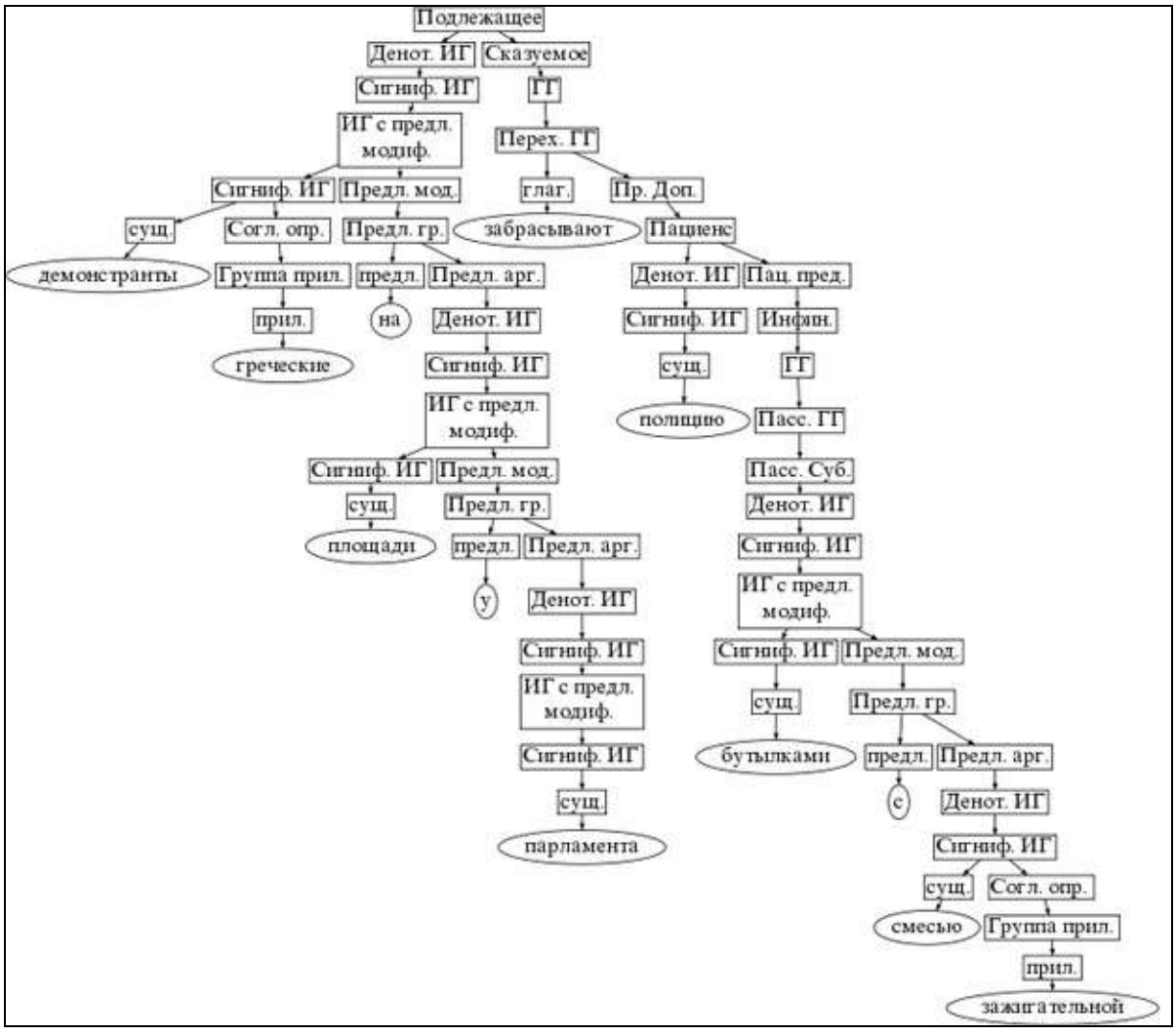

Рисунок 2

Точность автоматической рубрикации — это нормированная мера, определяющая для одного текста отношение количества корректно привязанных к нему рубрик к общему количеству рубрик, объективно релевантных данному тексту. Полнота автоматической рубрикации — это нормированная мера, определяющая для одного текста отношение количества корректно привязанных к нему рубрик к общему количеству рубрик, объективно релевантных данному тексту. Точность и полнота могут быть измерены для реакции системы автоматической рубрикации текстов на один текст, причем количество выданных результатов должно быть большим нуля. Формулы для определения данных параметров таковы:

$$Precision = \frac{|D_{rel} \cup D_{retr}|}{|D_{retr}|} \quad Recall = \frac{|D_{rel} \cup D_{retr}|}{|D_{rel}|},$$

где Precision — точность, Recall — полнота, D_{rel} — множество рубрик, релевантных тексту, D_{retr} — множество выданных рубрик.

Наиболее известной формулой для оценки эффективности системы автоматической рубрикации текстов является формула Ван Рисбергена (также известная как F-мера, или F_1-мера):

$$F_1 = \frac{2 * P * R}{P + R},$$ *где* ***P*** *— мера точности, а* ***R*** *— мера полноты.*

Проблему при определении параметров точности и полноты составляет множество D_{rel}: предполагается, что во всей совокупности рубрик можно выделить «строго» релевантные и «строго» нерелевантные анализируемому документу. Тем не менее, как отмечалось в [Добров 2010], релевантность рубрики документу — это мера, зависимая от множества параметров, каждый из которых может иметь разный вес.

При оценке статистической значимости различий между результатами работы системы автоматической рубрикации текстов и эталонной рубрикацией на основании T-критерия Стьюдента должно применяться сравнение выборочного среднего с заданным значением. Релевантными выданными рубриками можно считать рубрики, для которых T-критерий не показывает статистически значимых различий между выборочным средним значением релевантности, установленным экспертами, и значением релевантности, вычисленным системой автоматической рубрикации текстов. Предлагаемая методика оценки релевантности системы автоматической рубрикации текстов, в отличие от общепринятой, основывается на применении T-критерия Стьюдента для оценки точности, полноты и, соответственно, F-меры.

Таким образом, были выявлены основные особенности новостных сообщений, влияющие на эффективность их автоматической рубрикации. Была поставлена задача автоматической рубрикации новостных сообщений. Был предложен подход к созданию систем автоматической рубрикации новостных сообщений, основанный на использовании компьютерных методов комплексного лингвистического анализа текстов. Были проанализированы существующие методы оценки и сравнения эффективности работы систем автоматической рубрикации текстов и предложены некоторые уточнения к этим методам, на основании которых сформулирована уточненная методика оценки эффективности работы системы автоматической рубрикации текстов.

ЛИТЕРАТУРА

Агеев М. С., Добров Б. В., Лукашевич Н. В. Автоматическая рубрикация текстов: методы и проблемы // Уч. зап. КГУ. 2008. Т. 150, кн. 4.

Богатырёв М. Ю., Тюхтин В. В. Построение концептуальных графов как элементов семантической разметки текстов // Компьютерная лингвистика и интеллектуальные технологии : по матер. ежегодной Междунар. конф. «Диалог 2009» (Бекасово, 27—31 мая 2009 г.). — М.: РГГУ, 2009. Вып. 8 (15).

Добров А. В. Технологии интеллектуального поиска и способы оценки их эффективности // Структурная и прикладная лингвистика. — СПб.: Изд-во СПбГУ, 2010. Вып. 8.

Добросклонская Т. Г. Вопросы изучения медиатекстов (опыт исследования современной английской медиаречи). Изд. 2-е. — М.: Едиториал УРСС, 2005.

Мельчук И. А. Опыт теории лингвистических моделей «Смысл ↔ Текст». — М.: Наука, 1974.

Палагин А. В., Кривой С. Л., Петренко Н. Г. Концептуальные графы и семантические сети в системах обработки естественно-языковой информации // Математичні машини і системи. Киев, 2009. № 3.

Филиппов К. А. Лингвистика текста: курс лекций. — СПб: Изд-во СПбГУ, 2003.

Danes F. Functional sentence perspective and the organization of the text // Papers on functional sentence perspective. — Prague, 1974.

Kent A., Berry Madeline M., Luehrs Jr. Fr. U., Perry J.W. Machine literature searching VIII. Operational criteria for designing information retrieval systems // American Documentation. 1955. Vol. 6. Issue 2. P. 93—101.

Lewis D. Reuters-21578 text categorization test collection. Distribution 1.0. URL: http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt – 2004.

Sowa, John F. Conceptual Graphs for a Data Base Interface // IBM Journal of Research and Development. 1976. № 20 (4). P. 336—357.

Статью рекомендует к публикации д-р филол. наук, проф. А. П. Чудинов