А. В. Добров

АВТОМАТИЧЕСКАЯ РУБРИКАЦИЯ ТЕКСТОВ СРЕДСТВАМИ КОМПЛЕКСНОГО ЛИНГВИСТИЧЕСКОГО АНАЛИЗА

Аннотация. В статье рассматриваются современные методы автоматической рубрикации текстов и их недостатки. Предлагается комплексный лингвистический подход, позволяющий избежать этих недостатков. Описывается разработанный алгоритм автоматической обработки текстов и основанный на нем алгоритм выявления и ранжирования релевантных тексту рубрик.

Ключевые слова: автоматическая рубрикация текстов, неоднозначность, ложная корреляция, обработка текстов на естественном языке, семантика предложения.

A. V. Dobrov

AUTOMATIC TEXT CLASSIFICATION BASED ON COMPLEX LINGUISTIC ANALYSIS

Summary. In this article, modern methods of automatic text classification, and their shortcomings are discussed. A complex linguistic approach is proposed to avoid these shortcomings. The developed algorithm of automatic text processing, and the further algorithm to evaluate and rank topics, that are relevant to the text, are described.

Keywords: automatic text classification, ambiguity, false correlation, natural language processing, sentential semantics.

В современной лингвистической литературе не существует устоявшегося определения термина «автоматическая рубрикация». М. С. Агеев, Б. В. Добров и Н. В. Лукашевич определяют автоматическую рубрикацию как «отнесение порции информации к одной или нескольким категориям из ограниченного множества» (Агеев, Добров, Лукашевич 2008: 25), но не уточняют в рамках данного определения, о каком именно виде информации идет речь, что понимается под «категориями», и чем ограничено множество категорий. В то же

время эти авторы говорят далее не об отнесении информации к категориям, а об отнесении «текстов к рубрикам» (там же), которое они называют рубрикацией или классификацией текстов. По всей видимости, под «порциями информации» понимаются тексты, под множеством «категорий» — некое заранее определенное множество рубрик, к которым должны быть отнесены эти тексты, а под автоматической рубрикацией — отнесение машиной текстов к этим рубрикам. Следует различать рубрикацию и кластеризацию текстов. В результате кластеризации автоматически выстраиваются кластеры, к которым относятся тексты (Поддубный, Шевелев, Бормашов 2006; Васильев 2009), в то время как при автоматической рубрикации множество рубрик определяется заранее.

В некоторых работах для обозначения процедуры отнесения текстов к заранее заданным рубрикам используются также термины «тематическая классификация» (Инициативный проект… 2003: 6; Васильев 2008; 2010), «тематическое представление содержания документов» (Добров, Лукашевич 1996; Лукашевич, Добров 2009), «распознавание тематики текстов» (Белов, Волович 2007).

В современных системах автоматической рубрикации текстов (CAPT), как правило, используются формальные описания каждой рубрики, позволяющие относить или не относить к этой рубрике тот или иной текст. Такие описания называют «образами рубрик». Образы рубрик задаются в форме булевских формул, правил продукций или в иной форме.

В зависимости от того, каким образом строятся образы рубрик, можно выделить два класса методов автоматической рубрикации текстов — инженерные методы (методы, основанные на знаниях) и методы, основанные на машинном обучении. При применении инженерных методов образы рубрик создаются «инженерами по знаниям» (лингвистами и экспертами в различных предметных областях). При применении методов, основанных на машинном обучении, производится статистический анализ коллекции документов, предварительно распределенных по рубрикам вручную, на основании которого образы рубрик строятся автоматически.

В диссертационном исследовании М. С. Агеева (2005) подробно описаны различные методы, основанные на машинном обучении: метод Байеса, метод k-ближайших соседей, классификатор Rocchio, нейронные сети, деревья решений, построение булевых функций

(ПФА), а также метод опорных векторов (Support Vector Machines, SVM). Эти методы также описаны во многих работах семинара РОМИП (Агеев, Добров, Лукашевич, Сидоров, 2004; Агеев, Кураленок 2004; УИС РОССИЯ… 2008; Агеев, Добров, Лукашевич, Штернов 2008).

Методы автоматической рубрикации текстов, основанные на машинном обучении, вызывают интерес у исследователей, поскольку, на первый взгляд, позволяют понизить трудоемкость разработки САРТ, так как не требуют работы лингвистов и экспертов в различных предметных областях над составлением образов рубрик. Тем не менее, по данным исследования М. С. Агеева «… 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» (Агеев, Добров, Лукашевич 2008: 27). Для применения машинного обучения необходимо создать репрезентативную коллекцию текстов, распределенную по рубрикам в соответствии с принципами, единообразными для всей коллекции. Составление такой коллекции трудоемко, и, как правило, единообразие принципов отнесении текстов к рубрикам не гарантируется, что сказывается на достоверности выстраиваемых образов рубрик и на эффективности работы САРТ.

Правила рубрикации текстов, формализуемые в виде образов рубрик, часто основываются на наличии или отсутствии в текстах тех или иных лексических единиц. В простейшем случае правило отнесения текста к рубрике представляет собой дизъюнкцию наличия в тексте некоторых слов. В более сложном случае используются конъюнкции (требуется одновременное наличие двух или более слов) и отрицание (требуется отсутствие в тексте определенных слов). Такой подход приводит к ряду затруднений, связанных с морфологической и лексической неоднозначностью и с нехваткой информации о контексте словоупотреблений. Морфологическая неоднозначность может приводить к появлению ложных рубрик или нехватке правильных рубрик: например, при наличии в тексте словоформы «*стекло*» текст может быть ошибочно отнесен к рубрике «*строительные материалы*», в то время как в действительности в тексте употреблена форма единственного числа среднего рода прошедшего времени глагола «*стечь*». Многозначность лексических

единиц также может быть причиной некорректной рубрикации, когда слово употребляется в тексте не в том значении, на которое рассчитывал эксперт, составляя образ рубрики, или в котором оно было употреблено в текстах обучающей коллекции.

Проблемы лексической и морфологической неоднозначности достаточно существенны и сказываются на эффективности работы современных САРТ, но разрешимы, так как существуют различные способы снятия морфологической и даже лексической неоднозначности. Более существенна проблема «ложной корреляции», возникающая при использовании конъюнкций. Например, если при составлении образа рубрики «*Акции протеста*» используется конъюнкция наличия в тексте слов «*акция*» и «*протест*», то эта рубрика может быть присвоена тексту, в котором речь идет, например, о стоимости акций и о выражении акционерами протеста против ее повышения. Причина возникновения проблемы «ложной корреляции», как представляется, состоит в неверном подходе к созданию лингвистического обеспечения САРТ: для отнесения текста к рубрике «*Акции протеста*» необходимо присутствие в тексте не слов «*акция*» и «*протест*», а относящихся к этой рубрике понятий, выраженных синтаксически и семантически связанными языковыми единицами (например, «*акция протеста*», «*демонстрация протеста*», «*манифестация протеста*» и др.). Чтобы САРТ производила вычисления на основании понятий, а не слов, образ рубрики должен выстраиваться не в виде простых логических формул, требующих наличия или отсутствия в тексте лексических единиц, а в форме лингвистических моделей, отражающих семантические связи между понятиями. При этом для отнесения текста к рубрике необходимо производить не нормализацию присутствующих в тексте словоформ, а комплексный лингвистический анализ текста, позволяющий вычленить в тексте понятия, входящие в образ рубрики.

Принципы, используемые в современных САРТ при отнесении текстов к рубрикам, не дают возможности произвести ранжирование этих рубрик. Текст, в котором речь идет об акции протеста, но в одном параграфе упоминается экономическая реформа, в большей мере относится к акциям протеста, чем к экономической реформе; тем не менее, если выражения «*акция протеста*» и «*экономическая реформа*» встречаются в тексте одинаковое количество раз, то САРТ не может присвоить этим рубрикам разный вес. Комплекс-

ный лингвистический анализ текста дает возможность решить также и эту проблему путем выявления тематической (коммуникативной) структуры дискурса.

Системы, производящие комплексную автоматическую обработку текстов, называют лингвопроцессорами. Лингвопроцессор должен быть способен вычислять смысловое содержание текста, имея на входе только сам текст и полное описание того языка, на котором написан этот текст. Модели содержания текста, выстраиваемые лингвопроцессором, должны быть совместимыми с правилами рубрикации, содержащимися в образах рубрик.

Комплексный лингвистический анализ текста, производящийся лингвопроцессором, включает в себя выявление языковых единиц различных языковых уровней. Н. Н. Леонтьева называет такой анализ текста «автоматическим пониманием текста» (АПТ), подчеркивая, что «в отличие от многих других систем АОТ системы АПТ обладают максимальным набором лингвистических компонентов — это полные системы» (Леонтьева 2006: 10). Традиционно выделяют три вида анализа, производимого при комплексной автоматической обработки текста, — морфологический, синтаксический и семантический. В результате морфологического анализа в тексте выявляются словоформы, которые снабжаются грамматической информацией и соотносятся с лексемами. В результате синтаксического анализа на основании полученной грамматической информации о каждой словоформе и их линейного порядка выявляются словосочетания, предложения и сверхфразовые единства, состоящие из этих словоформ. В результате семантического анализа на основании семантики синтаксических связей между частями выявленных в процессе синтаксического анализа составляющих и на основании правил семантической сочетаемости значений лексем вычисляется совокупность понятий и семантических отношений, образующих смысловое содержание текста.

В ранних моделях автоматической обработки текстов, в частности, в модели И. А. Мельчука «Смысл ⇔ Текст» (Мельчук 1974) и в моделях его последователей (напр., (Леонтьева 2006)) предполагалась последовательная работа морфологического, синтаксического и семантического компонентов. На первый взгляд, такой подход позволяет упростить процесс обработки текста путем разбиения его на этапы с очевидными правилами перехода между этапами. Тем не

менее практические применения таких моделей (ср., напр., система «ЭТАП» (Ю. Д. Апресян и др.) показали, что на каждом этапе автоматической обработки текста возникает множество в равной мере допустимых версий анализа, и если на этапе морфологического анализа строится в достаточной мере ограниченное количество версий, то уже на этапе синтаксического анализа возникает практически непреодолимая проблема «комбинаторного взрыва», возникающая в результате многократной синтаксической неоднозначности в пределах одного предложения. Проблема «комбинаторного взрыва» состоит в том, что в процессе синтаксического анализа лингвопроцессор должен выполнить перебор всех возможных комбинаций различных интерпретаций компонентов этого предложения, от словоформ до полностью распространенных словосочетаний.

Синтаксическая неоднозначность может возникать и как следствие морфологической неоднозначности, и безотносительно к ней. Например, предложение «*Эти типы стали есть в цехе*» характеризуется синтаксической неоднозначностью вследствие морфологической неоднозначности словоформы «*стали*», которая может быть отнесена как к имени существительному «*сталь*», так и к глаголу «*стать*». Синтаксическая неоднозначность, не обусловленная морфологической неоднозначностью, составляет более существенную проблему. Например, предложение «*Миронов встретился с Лужковым в Москве*» также характеризуется синтаксической неоднозначностью: фрагмент «*в Москве*» может быть распознан не только как обстоятельство места, относящееся к сказуемому «*встретился*», но и как предложное определение, относящееся к дополнению «*с Лужковым*», так как с формально-грамматической точки зрения это предложение ничем не отличается от, например, предложения «*Иванов встретился с мальчиком в шляпе*». Такие виды синтаксической неоднозначности носят системный характер, поскольку не зависят от конкретного лексического наполнения синтаксической структуры.

Синтаксическая неоднозначность может быть обусловлена также возможностью восстановления эллипсиса в отдельных частях предложения. Такая ситуация наблюдается, например, в следующих двух предложениях: «*Мусульманские женщины носят паранджу. Иудейские мужчины — кипы*». С формальной точки зрения, второе предложение можно интерпретировать как дефиницию: фрагмент «*кипы*» может быть составным именным сказуемым, и предложение

может означать: “Иудейские мужчины — это кипы”. Корректная интерпретация синтаксической структуры такого предложения может быть построена только с учетом возможности эллипсиса сказуемого «*носят*». Для этого при автоматическом синтаксическом анализе должны выстраиваться не только те синтаксические структуры анализируемого предложения, которые напрямую соответствуют представленным в тексте словоформам, но и те, которые содержат в себе эллиптированные элементы, не представленные в анализируемом предложении. Так как синтаксические анализаторы, как правило, обрабатывают каждое предложение в отдельности, а не несколько предложений одновременно, возникает необходимость предполагать возможность эллипсиса во всех допускающих эллипсис узлах синтаксической структуры. Проблема осложняется тем, что при эллипсисе кореферентного антецедента может не быть: например, в вопросе «*Вам кого?*», по всей вероятности, эллиптированы, по крайней мере, например, словоформы «*следует*» и «*позвать*», никак не представленные в предыдущих предложениях, поскольку предыдущих предложений нет.

Указанные факторы, приводящие к синтаксической неоднозначности предложений, действуют одновременно и многократно практически в любом предложении, поэтому количество комбинаций, перебор которых должен быть выполнен лингвопроцессором при синтаксическом анализе, крайне велико. Как отмечает Н. Н. Леонтьева, «получить хорошие результаты СинАн для всех предложений естественного, непрепарированного массива текстов оказывается практически невыполнимой или безмерно сложной задачей. Причин тому много. Это и негладкость построения многих реальных предложений <…> и, самое главное, очень большая локальная неоднозначность» (Леонтьева 2006: 79,80).

Существуют различные подходы к решению проблемы синтаксической неоднозначности. Как отмечает И. С. Евдокимова, «с точки зрения цели синтаксического анализа можно выделить два основных подхода: одноцелевой и многоцелевой. При первом подходе для фразы требуется построить одно синтаксическое представление, этот подход характерен для первых алгоритмов синтаксического анализа, когда считалось, что синтаксических средств достаточно для того, чтобы обеспечить правильный анализ фразы, хотя бы для большинства фраз. При втором подходе для фразы требуется полу-

чить все те синтаксические представления, которые удовлетворяют определенным соглашениям …» (Евдокимова 2006: 75).

Следует отметить, что помимо одноцелевых и многоцелевых методов синтаксического анализа существуют также комбинированные методы, сочетающие в себе приемы обоих методов, в частности — метод фильтров, при котором различные версии анализа не разделяются, а, напротив, объединяются в единый недревовидный граф. Основные идеи метода фильтров были сформулированы в работах И. Лесерфа и развиты рядом других исследователей, в частности Л. Н. Иорданской, Ю. Д. Апресяном.

Метод фильтров способствует экономии машинных ресурсов без потери лингвистической корректности результатов анализа. Вместо построения нескольких независимых синтаксических структур выстраивается единая структура, содержащая в себе как однозначные, так и неоднозначные связи. При этом исчезает необходимость многократного копирования частей этой структуры, общих для разных версий анализа. Лингвистическая корректность этого метода обусловлена тем, что фильтры могут быть и синтаксическими, и семантическими, так как применение фильтров не составляет части непосредственно синтаксического анализа.

Предлагаемый в статье подход к организации работы лингвопроцессора основан на одновременном выполнении морфологического, синтаксического и семантического анализа текста, предполагающем фильтрацию части версий на каждом из уровней при выявлении невозможности построения версий более высокого уровня. Так, частичное снятие морфологической неоднозначности может производиться уже тогда, когда обнаруживается невозможность построения синтаксической связи между словоформами. На каждом шаге синтаксического связывания должен выполняться семантический анализ полученной синтаксической связи, в процессе которого фильтруется часть версий синтаксического анализа. Например, при разборе предложения «*Демонстранты забрасывают полицию бутылками с зажигательной смесью*» в одной из версий синтаксического анализа компонент «*с зажигательной смесью*» может рассматриваться как обстоятельство при сказуемом «*забрасывают*», означающее совместность действия. Семантический анализ показывает некорректность такой синтаксической связи, так как совместность действия предполагает отнесенность предмета, обозначаемого об-

стоятельством, к живым существам, но выражение «*бутылки с зажигательной смесью*» не удовлетворяет этому требованию ни в одном из своих значений.

Такой подход позволяет существенно повысить производительность лингвопроцессора и фактически преодолеть проблему комбинаторного взрыва. При иной организации работы лингвопроцессора проверка семантических ограничений осуществлялась бы на более поздних этапах, и некорректные с точки зрения семантики гипотезы порождали бы множество других, излишних гипотез. Выбор такого способа организации работы лингвопроцессора обусловлен, однако, не только соображениями производительности, но и тем, что семантика, как указывает А. С. Герд, «*пронизывает все уровни языка* и тем самым не представляет собой отдельного уровня» (Герд 1996: 9). Именно поэтому производить морфологический или синтаксический анализ текста в отрыве от семантического анализа, как представляется, не только нерационально, но и не вполне корректно с лингвистической точки зрения.

Изложенный подход к организации работы лингвопроцессора дает возможность производить автоматическую обработку текста в соответствии со следующим абстрактным алгоритмом.

1. Во входном потоке выделяются атомарные единицы.

2. Для каждой выделенной единицы производится попытка семантического анализа.

3. Для каждой пары соседних выделенных единиц, для которых в п. 2 был получен некий результат, производится попытка объединения этой пары в единую составляющую. В случае успеха полученная составляющая считается новой выделенной единицей, для которой повторяются пункты 2–3.

Этот алгоритм в одинаковой мере применим как к морфологическому, так и к синтаксическому анализу. На уровне морфологического анализа атомарными единицами выступают морфемы, а сложными составляющими — словоформы и входящие в их состав словоизменительные и словообразовательные основы. В процессе синтаксического анализа словоформы объединяются в составляющие более высоких уровней — словосочетания, предложения и сверхфразовые единства.

Объединение двух единиц в единую составляющую, выполняющееся в п. 3, производится путем перебора нециклических маршру-

тов, связывающих классы этих единиц в грамматике составляющих, представленной в виде графа, узлы которого соответствуют классам составляющих, а дуги связывают каждый класс составляющих с классами их возможных дочерних составляющих. Класс общей составляющей располагается в той точке маршрута, в которой изменяется его направление: так как структуры составляющих древовидны, маршрут должен проходить от первой связываемой составляющей по цепочке родительских составляющих (вверх) до тех пор, пока не найдется такая составляющая, от которой существует возможное продолжение этого маршрута по цепочке дочерних составляющих (вниз) до второй связываемой составляющей.

Семантический анализ выделенных единиц, выполняющийся в п. 2, функционирует следующим образом: если анализируемая единица атомарная или входит в множество «идиоматических» единиц (т. е. таких единиц, значение которых невозможно вычислить из значений их частей), то совокупность значений этой единицы извлекается из словаря; если же анализируемая единица состоит из более простых составляющих и не идиоматическая, то ее значение вычисляется путем попытки установить между значениями этих составляющих семантическое отношение, соответствующее типу и направлению синтаксической, словообразовательной или формообразовательной связи.

Этот алгоритм реализован в виде программы на языке C, распространяющейся в исходных кодах на некоммерческой основе под свободной лицензией GNU GPL и входящей в дистрибутивы операционной системы «НауЛинукс». Разработанный лингвопроцессор производит построение возможных семантических представлений предложений и сверхфразовых единств анализируемого текста. В процессе семантического анализа выполняется также попытка разрешения анафорических связей с учетом интерпретации линейного порядка составляющих с точки зрения актуального членения. Анафорическая связь предполагается для каждой пары элементов семантического представления двух линейно соседних простых предложений, если второй элемент пары тематический с точки зрения актуального членения, и между элементами существуют отношения синонимии или гиперонимии. Указанное правило распространяется не только на именные, но и на глагольные группы. Кроме того, предполагаются анафорические связи между местоимениями

и соответствующими их значениям членами предшествующего (для рефлексивного местоимения — того же) простого предложения.

Устанавливаемые в соответствии с указанными правилами анафорические связи образуют сеть тематических прогрессий, позволяющую производить автоматическую рубрикацию анализируемого текста и ранжирование приписываемых ему рубрик. Ранжирование рубрик производится на основании вычисления и суммирования рангов понятий (элементов семантического представления), образующих образы этих рубрик. Ранги понятий вычисляются в соответствии со следующими правилами:

1) ранг понятия, выраженного синтаксической составляющей, не связанной ни напрямую, ни опосредованно никакими анафорическими связями с другими составляющими, равен единице;

2) понятия, выраженные тематическими составляющими, связанными с другими составляющими напрямую или опосредованно анафорическими связями, обладают одинаковым рангом;

3) понятия, выраженные рематическими составляющими, связанными с другими составляющими опосредованно анафорическими связями, обладают рангом, на единицу большим, чем ранг антецедента.

Чем выше ранг понятия, тем в меньшей мере рубрики, к которым относится это понятие, характеризуют анализируемый текст. Поэтому релевантность понятия тексту целесообразно считать обратно пропорциональной его рангу. Релевантность рубрики можно определить как отношение суммы релевантностей тексту понятий, входящих в ее образ, к сумме релевантностей понятий, входящих в образ любой рубрики, относящейся к данному тексту:

$$R_{rub} = \frac{\sum\limits_{c \in Img(rub)} R_c}{\sum\limits_{\exists r : c \in Img(r)} R_c},$$

где *rub* — рубрика, *Img*(*r*) — образ рубрики *r*, R_c — релевантность *c*.

Согласно этой формуле, релевантность рубрики R_c не превышает 1 и не может быть отрицательной; сумма релевантностей всех присвоенных тексту рубрик составляет 1.

Литература

Агеев М. С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов: дис. … канд. физ.-мат. наук. М., 2005.

Агеев М. С., Добров Б. В., Лукашевич Н. В. Автоматическая рубрикация текстов: методы и проблемы // Учен. зап. Казанск. гос. ун-та. Т. 150, кн. 4.

Агеев М. С., Добров Б. В., Лукашевич Н. В., Сидоров А. В. Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // Российский семинар по оценке методов информационного поиска (РОМИП 2004): Семинар в рамках Всерос. науч. конф. RCDL'2004. 1 окт. 2004 г. Пущино, 2004.

Агеев М. С., Добров Б. В., Лукашевич Н. В., Штернов С. В. УИС РОССИЯ в РОМИП 2008: поиск и классификация нормативных документов // Российский семинар по оценке методов информационного поиска (Труды РОМИП 2007–2008): Семинар в рамках Всерос. науч. конф. RCDL'2008. 9 окт. 2008 г., Дубна. СПб., 2008.

Агеев М. С., Кураленок И. Е. Официальные метрики РОМИП'2004 // Российский семинар по оценке методов информационного Поиска (РОМИП 2004). Пущино, 2004.

Белов А. А., Волович М. М. Автоматическое распознавание тематики сверхкоротких текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды междунар. конф. «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) / под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. М., 2007. С. 35–37.

Инициативный проект российского семинара по оценке методов информационного поиска (РОМИП) / П. И. Браславский, М. В. Губин, Б. В. Добров, В. Ю. Добрынин, И. Е. Кураленок, И. С. Некрестьянов, Е. Ю. Павлова, И. В. Сегалович // Компьютерная лингвистика и интеллектуальные технологии: труды Междунар. конф. Диалог–2003. Протвино. 11–16 июня 2003 г. / под ред. И. М. Кобозевой, Н. И. Лауфер, В. П. Селегея. М., 2003.

Васильев В. Г. Комплексная технология автоматической классификации текстов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог» (Бекасово, 4–8 июня 2008 г.). М., 2008. Вып. 7 (14).

Васильев В. Г. Выделение фрагментов в текстах при классифика-

ции // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Междунар. конф. «Диалог–2009» (Бекасово, 27–31 мая 2009 г.). М., 2009. Вып. 8 (15).

Васильев В. Г. Обучение классификаторов на основе выделения фрагментов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог» (Бекасово, 26–30 мая 2010 г.). М., 2010. Вып. 9 (16).

Герд А. С. Предмет и основные направления прикладной лингвистики // Прикладное языкознание: учебник. СПб., 1996.

Добров Б. В., Лукашевич Н. В. Построение и использование тематического представления содержания документов // V национальная конференция с международным участием «Искусственный интеллект-96». Казань, 1996. Т. 1. С. 130–134.

Евдокимова И. С. Естественно-языковые системы. Улан-Удэ, 2006.

Леонтьева Н. Н. Автоматическое понимание текстов. Системы, модели, ресурсы. М., 2006/

Лукашевич Н. В., Добров Б. В. Автоматическое аннотирование новостных кластеров на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Междунар. конф. «Диалог–2009» (Бекасово, 27–31 мая 2009 г.). М., 2009. Вып. 8 (15).

Мельчук И. А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., 1974.

Поддубный В. В., Шевелев О. Г., Бормашов Д. А. Сравнение качества подходов к кластеризации текстов на основе гипергеометрического критерия. // Вестник Томск. гос. ун-та. 2006. N 293.

УИС РОССИЯ в РОМИП'2007: поиск и классификация / М. С. Агеев, Б. В. Добров, П. В. Красильников, Н. В. Лукашевич, А. М. Павлов, А. В. Сидоров, С. В. Штернов // Российский семинар по оценке методов информационного поиска (Труды РОМИП 2007–2008): Семинар в рамках Всерос. науч. конф. RCDL'2007. 18 окт. 2007 г., Переславль-Залесский. СПб., 2008.